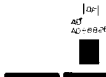
AD-A086 826 NAVAL OCEAN SYSTEMS CENTER SAN DIEGO CA F/G 12/1
TWO'S-COMPLEMENT FIXED-POINT MULTIPLICATION ERRORS - THEORY.(U)
APR 80 L P MULCAHY
UNCLASSIFIED NOSC/TR-538 NL



END
DATE FILMED
8-80
DTIC

NOSC

NOSC TR 538

NOSC TR 538

Technical Report 538

TWO'S-COMPLEMENT FIXED-POINT MULTIPLICATION ERRORS — THEORY

LP Mulcahy

1 April 1980

Final Report: May 1975 — September 1979

Approved for public release; distribution unlimited.

NAVAL OCEAN SYSTEMS CENTER
SAN DIEGO, CALIFORNIA 92152

NAVAL OCEAN SYSTEMS CENTER, SAN DIEGO, CA 92152

AN ACTIVITY OF THE NAVAL MATERIAL COMMAND

SL GUILLE, CAPT, USN
Commander

HL BLOOD
Technical Director

ADMINISTRATIVE INFORMATION

This work was an unfunded outgrowth of work supported by Independent Exploratory Development funds at NOSC. Work reported herein was performed from May 1975 through September 1979.

Reviewed by
R. W. Larsen, Head
Systems Validation and Support Division

Under authority of
E. B. Tunstall, Head
Ocean Surveillance Systems Department

ACKNOWLEDGMENT

The author wishes to thank Messrs. J. W. Bond and J. M. Speiser of the Naval Ocean Systems Center for their helpful comments, suggestions, and reviews of technical accuracy.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: NOSC Technical Report 538 (TR 538)	2. GOVT ACCESSION NO.: AD-A086 826
	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)*: TWO'S-COMPLEMENT FIXED-POINT MULTIPLICATION ERRORS – THEORY	5. TYPE OF REPORT & PERIOD COVERED: Final Report May 1975–September 1979
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): L. P. Mulcahy	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Naval Ocean Systems Center, San Diego, CA 92152	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS	12. REPORT DATE: 1 April 1980
	13. NUMBER OF PAGES: 57
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*: 12 59	15. SECURITY CLASS. *(of this report)*: Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT *(of this Report)*: Approved for public release; distribution unlimited. 14 NOSC/TR-538	
17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*	
18. SUPPLEMENTARY NOTES	
19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*: digital filters; roundoff; chopping; two's-complement; multiplication errors; fixed-point	

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Analytic forms of a variety of two's-complement multiplication error statistics are derived. An FIR filter structure is used to define the individual error statistics which are used in the calculation of the filter output variance and spectrum. The approach used in deriving the statistics is to show that a regularity exists in the structure of the errors as the multiplier input is stepped through a series of consecutive values. The error structure is a function of coefficient value, number representation, and word lengths. This regular structure allows the use of the Poisson Summation Formula in deriving error statistics based on multiplier inputs obtained from quantized random variables. The error statistics are categorized according to families of coefficient values denoted by the parameter ν. This parameter is the effective word length of the error. Specific forms of the statistics are given for Gaussian quantizer inputs.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

157

SUMMARY

PROBLEM

Extend the theory of computation error generation in digital filters. Specifically, consider two's-complement fixed-point multiplication for digitized Gaussian signal inputs.

RESULTS

Deterministic properties of round-off and chopping were examined for two's-complement fixed-point multiplication errors. It was shown that a regularity exists in the structure of the errors as the multiplier input is stepped through a series of consecutive values. The error structure is a function of coefficient value, number representation, and word lengths. The error properties are categorized according to families of coefficient values denoted by the parameter ν. This parameter is the effective word length of the error. This regular structure allowed the use of the Poisson Summation Formula in deriving error statistics based on multiplier inputs obtained from quantized random variables.

A finite impulse response filter structure was used to define the individual error statistics which are used in the calculation of the filter output variance and spectrum. The error statistics which were derived are:

(1) error probabilities (univariate and bivariate),

(2) mean error,

(3) second moment of the error,

(4) the cross-correlation between a multiplier input and the error for the same or a different multiplier,

(5) the autocorrelation of the error for one multiplier, and

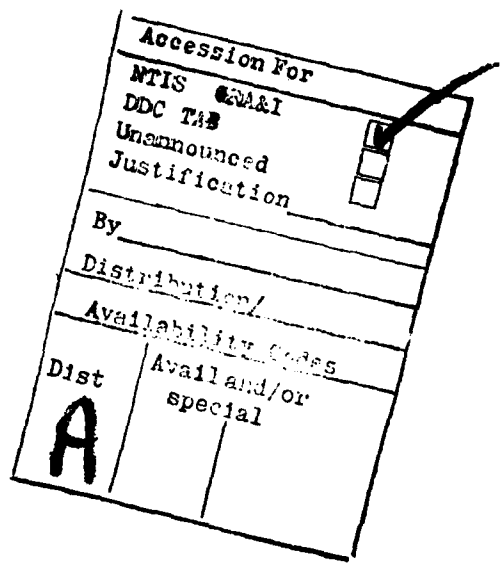
(6) the cross-correlation between the errors for two multipliers.

The analytical form of each statistic was shown to consist of two parts, an asymptotic part and a decreasing part. The asymptotic part, which is dependent on ν, is well behaved and finite for finite filter input mean. The asymptotic part was shown to be independent of the quantizer input probability density function shape. The only asymptotic part which depends on quantizer input statistics appears in the cross-correlation between a multiplier input and the error for the same or a different multiplier. The quantizer input mean appears there. The decreasing part is in the form of a summation which varies as a function of the input standard deviation. It has the property that it decreases to zero in the limit as the input standard deviation is increased. Specific forms of the decreasing part of the statistics were given for Gaussian quantizer input probability density functions.

RECOMMENDATIONS

Examine the error statistics presented in this report relative to assumptions and engineering guidelines used presently. Show where present guidelines are adequate and where modifications are needed.

CONTENTS



I. INTRODUCTION

The usual approach to the statistical analysis of the effects of two's-complement (TC) roundoff errors in digital filters makes use of the following argument. Since multiplication errors are similar to analog signal quantization errors, the same statistical results are assumed to hold. The multiplication error can be represented as a white noise which is zero-mean, which is uniformly distributed over the magnitude of the least significant bit (l.s.b.) after rounding, and which has zero cross-correlation with the multiplier input.

Chopping errors are also assumed uniformly distributed, but they have a non-zero mean for the TC number representation [1]-[2].

Recent results [3] presented a more accurate model of the error generation process that is valid for "large" standard deviation of the multiplier input. Error statistics were derived showing their dependence on number representation, coefficient value, chopping or rounding scheme, and certain statistics of the multiplier input. Statistics for other than large standard deviation were not obtained.

This report derives analytic expressions for pertinent statistics of TC multiplication errors for the case of arbitrary standard deviation of the multiplier input. It starts first by examining the form of the finite impulse response (FIR) filter and determining which statistics of the error are important. Next, the deterministic TC error properties are derived. The analog signal quantization process is discussed and certain useful formulas are presented through use of the Poisson Summation Formula (hereafter abbreviated as PSF)[4]. The desired error statistics are then derived, also through use of the PSF. Specific forms of the statistics are presented for the case of a Gaussian quantizer input.

II. THE FILTER MODEL

This section is presented as motivation for the choice of error statistics which are subsequently derived in this paper.

A block diagram representation of an FIR filter is shown in Fig. 1 in the form of a tapped delay line. The number of taps employed is equal to U and the time interval between consecutive data samples is equal to T. The taps are weighted by the filter coefficient values $\{a_h; h = 0,1, \ldots, U-1\}$. These are the quantized values which have been determined through some design procedure. The filter input data are the stationary random sequence values $\{x(iT); i = 0,\pm1,\pm2, \ldots\}$. This sequence is obtained by sampling and quantization (with roundoff) of the analog waveform $\tilde{x}(t)$. The filter incorporates the multiplication error as the error sequence $\{\epsilon_h([i-h]T); i = 0,\pm1,\pm2, \ldots\}$. The index $h = 0, 1, \ldots, U-1$ identifies the tap the error sequence is associated with. As shown in [3], the error sequence values are deterministically related to the corresponding multiplier input values. Hence, the time index for the error sequence is written as shown to avoid confusion later on when evaluating the error statistics associated with particular multiplier inputs. The filter output data are the sequence values $\{y(iT); i = 0, \pm1, \pm2, \ldots\}$. The number values associated with x, a, and y are assumed exactly representable by binary words of lengths K, L and M bits, respectively. A leading binary point is also assumed. These lengths are exclusive of sign bits. The input/output relation for this filter is

$$y(iT) = b(iT) + w(iT) \tag{1}$$

where

$$b(iT) = \sum_{h=0}^{U-1} a_h x([i-h]T) \tag{2}$$

$$w(iT) = \sum_{h=0}^{U-1} \epsilon_h([i-h]T) \, . \tag{3}$$

The analysis presented in this paper does not depend only on the use of sampling and quantization of an analog waveform. The filter input can come from the output of another digital filter for example. It turns out that the statistical results can be written in terms of the filter input alone or in terms of the quantizer input. The latter is more interesting in that specific forms for the Gaussian case can then be applied to the analysis.

From [3], multiplier overflow occurs only for certain coefficient values and for a small number of values of x, if at all, at one or the other end point of the range of x (i.e., at or close to plus or minus 1.0, depending on the value of the coefficient). (Values of x and a for which overflow occurs are defined in this paper.) The sums represented by b and w can also result in filter overflow. In the analysis it will be assumed that the probability of occurrence of either kind of overflow is so small as to be negligible. Practically, this can be accomplished by reducing the variance of the input sequence or by increasing the input and output word lengths K and M.

The sequence {b(iT)} in (2) is the desired filter output. The sequence {w(iT)} represents additive noise which is deterministically related to the input sequence. However, the usual statistical descriptions can be employed. These are the sequence mean, the variance, and the power spectral density.

The mean of y(iT) is given by $\mu_y = E[y(iT)]$ where E denotes expected value. It is computed from (1) as

$$\mu_y = \mu_x \sum_{h=0}^{U-1} a_h + \sum_{h=0}^{U-1} \mu_{\epsilon_h} \quad . \tag{4}$$

The variance of y(iT) is given by σ_y^2 where

$$\sigma_y^2 = E[y^2] - \mu_y^2 \tag{5}$$

$$= \sigma_b^2 + \Delta_\sigma \quad . \tag{6}$$

The desired variance is given by σ_b^2 where

$$\sigma_b^2 = \sum_{g=0}^{U-1} \sum_{h=0}^{U-1} [a_g a_h C_x([g-h]T)] \quad . \tag{7}$$

The term Δ_σ represents the change imposed on the desired variance by the multiplication errors. It can take on negative values and is given by the following expression:

$$\Delta_\sigma = \sum_{g=0}^{U-1} \sum_{h=0}^{U-1} [2a_g C_{x_g \epsilon_h}([g-h]T) + C_{\epsilon_g \epsilon_h}([g-h]T)] \quad . \tag{8}$$

The autocovariance $C_x(\cdot)$ and cross-covariances $C_{x_g \epsilon_h}(\cdot)$ and $C_{\epsilon_g \epsilon_h}(\cdot)$ are defined in terms of their corresponding auto- and cross-correlations and mean values as follows:

$$C_x(\cdot) = R_x(\cdot) - \mu_x^2 \, , \tag{9}$$

$$C_{x_g \epsilon_h}(\cdot) = R_{x_g \epsilon_h}(\cdot) - \mu_x \mu_{\epsilon_h} \, , \tag{10}$$

$$C_{\epsilon_g \epsilon_h}(\cdot) = R_{\epsilon_g \epsilon_h}(\cdot) - \mu_{\epsilon_g} \mu_{\epsilon_h} \quad . \tag{11}$$

The power spectral density of y(iT) is given by $S_y(\omega)$ where

$$S_y(\omega) = \sum_{\eta=-\infty}^{\infty} R_y(\eta T) \exp(-j\omega\eta T) \tag{12}$$

and $R_y(\eta T)$ is the autocorrelation of y(iT). The autocorrelation in turn can be written as

$$R_y(\eta T) = R_b(\eta T) + \Delta_R(\eta T) \tag{13}$$

where $\eta = 0, \pm 1, \pm 2, \ldots$.

The desired autocorrelation is given by $R_b(\eta T)$ where

$$R_b(\eta T) = \sum_{g=0}^{U-1} \sum_{h=0}^{U-1} a_g a_h R_x([g + \eta - h]T) \ . \tag{14}$$

The term $\Delta_R(\eta T)$ represents the change in the desired autocorrelation and, hence, the spectrum. It is given by the following expression:

$$\Delta_R(\eta T) = \sum_{g=0}^{U-1} \sum_{h=0}^{U-1} \left[2a_g R_{x_g \epsilon_h}([g+\eta-h]T) + R_{\epsilon_g \epsilon_h}([g+\eta-h]T) \right] \ . \tag{15}$$

Thus, the spectrum in turn can be written as

$$S_y(\omega) = S_b(\omega) + \Delta_S(\omega) \tag{16}$$

where

$$S_b(\omega) = \sum_{\eta=-\infty}^{\infty} R_b(\eta T) \exp(-j\omega\eta T) \tag{17}$$

and

$$\Delta_S(\omega) = \sum_{\eta=-\infty}^{\infty} \Delta_R(\eta T) \exp(-j\omega\eta T) \ . \tag{18}$$

The time delay argument $[(g+\eta-h)T]$ comes about because the same data sequence x(iT) is used as input to all multipliers. Let $x_g(iT)$ and $x_h(iT)$ be the inputs to multipliers g and h. However,

$$x_g(iT) = x([i-g]T) \tag{19}$$

and

$$x_h(iT) = x([i-h]T) \ . \tag{20}$$

Then, for example,

$$R_{x_g x_h}(\eta T) = E[x_g(iT)x_h([i+\eta]T)]$$

$$= R_x([g+\eta-h]T) \ . \tag{21}$$

Since the multiplication errors and multiplier outputs are deterministically related to the inputs, it seems most natural to write the second order statistics, $R_{x_g \epsilon_h}$ and $R_{\epsilon_g \epsilon_h}$, in terms of the time delays imposed on the input sequence x(iT). The approach to the derivation of the equations to be used for these statistics depends on whether $\rho = 1.0$ or $|\rho| < 1.0$ for the quantizer input sequences used.* When g = h in the time delay argument, $\rho = 1.0$ for $\eta = 0$ and $|\rho| < 1.0$ for $\eta \neq 0$. If $g \neq h$, then $\rho = 1.0$ for $h-g = \eta$ and $|\rho| < 1.0$ for $h-g \neq \eta$. Note that when g = h only one coefficient value is involved. Then

$$R_{x_g \epsilon_h}(\cdot) = R_{x_g \epsilon_g}(\eta T) \tag{22}$$

and

$$R_{\epsilon_g \epsilon_h}(\cdot) = R_{\epsilon_g}(\eta T) \ . \tag{23}$$

When $g \neq h$,

$$R_{x_g \epsilon_h}(\cdot) = R_{x_g \epsilon_h}([g+\eta-h]T) \tag{24}$$

and

$$R_{\epsilon_g \epsilon_h}(\cdot) = R_{\epsilon_g \epsilon_h}([g+\eta-h]T) \ . \tag{25}$$

Computationally, (24) is no different than (22) since only one coefficient value is involved. However, (25) is not usually the same as (23) since two coefficient values are involved. If the two coefficients are different, two different transformations from x to ϵ exist. Then (25) must be used. If the coefficients are equal, (23) may be used. Note that two separate, equal coefficient values lead to the curious situation where the cross-correlation of both error sequences is equal to the autocorrelation of either error sequence alone except for a constant difference in the time delay argument.

* $\rho = \rho_{\underset{\sim}{x}}$ is defined in Section IV.

The statistics necessary for the evaluation of the covariances are then the following: the filter input mean μ_x, the filter input autocorrelation $R_x(\cdot)$, the mean multiplication error μ_ϵ, the error autocorrelation $R_\epsilon(\cdot)$, the error cross-correlation $R_{\epsilon_g \epsilon_h}(\cdot)$, and the cross-correlation between error and multiplier input $R_{x_g \epsilon_h}(\cdot)$. These are the quantities for which analytical expressions will be presented.

III. DETERMINISTIC ERROR PROPERTIES

COMMENTS ON NOTATION

Multiplier word lengths are defined as follows: input x, K bits plus sign; coefficient a, L bits plus sign; result of roundoff or chopping y, M bits plus sign; and N bits which are eliminated through roundoff or chopping. The lengths K, L, M, and N are arbitrary constant positive integers with the only restriction that K+L = M+N. The relationships among the various word lengths are shown in the diagram in Fig. 2. The coefficient dependent parameters ν and δ are also shown.

The following definitions will be used in this report. Let k be any positive or negative integer. Then $[k]_{2^N}$ is the set of non-negative integers such that $0 \leqslant [k]_{2^N} \leqslant (2^N-1)$ where $[k]_{2^N}$ is related to k through the equivalence relation

$$[k]_{2^N} \equiv k \bmod 2^N. \tag{26}$$

The indicator function $I_{2^N}[k]$ has the property that

$$I_{2^N}[k] = \begin{cases} 1 \text{ for } 2^{N-1} \leqslant [k]_{2^N} \leqslant 2^N - 1 , \\ \\ 0 \text{ for } 0 \leqslant [k]_{2^N} \leqslant 2^{N-1} - 1 . \end{cases} \tag{27}$$

NUMBER REPRESENTATION

Consider the number value b and its (M+N)-bit binary fixed-point machine representation b*. The machine representation is treated as a positive number and is defined as

$$b^* = b_0 . b_1 b_2 \cdots b_{M+N} = \sum_{i=0}^{M+N} b_i 2^{-i} . \tag{28}$$

The following relationships exist between b and b*. For non-negative numbers $b \geqslant 0$,

$$b^* = b . \tag{29}$$

For negative numbers $b < 0$,

$$b^* = 2 + b . \tag{30}$$

The binary number b_0 is called the sign bit (i.e., $b_0 = (1\text{-sgn } b)/2$ for $b \neq 0$ and $b_0 = 0$ for $b = 0$).

Let b be written as

$$b = u2^{-M-N} \tag{31}$$

where the integer u can take on the values

$$u = 0, 1, \ldots, 2^{M+N} - 1 \tag{32}$$

for all non-negative numbers $u \geq 0$, and can take on the values

$$u = -2^{M+N}, \ldots, -2, -1 \tag{33}$$

for negative numbers $u < 0$. The remainder r is a non-negative number which is defined through the following. Let

$$b^* = \sum_{i=0}^{M} b_i 2^{-i} + r^* 2^{-M} \tag{34}$$

where

$$\begin{aligned} r^* &= r \\ &= 0 . b_{M+1} b_{M+2} \cdots b_{M+N} \\ &= 2^{-N} [2^{M+N} b^*]_{2^N} . \end{aligned} \tag{35}$$

Let y^* be the machine representation of the result of roundoff or chopping of b^*. Then

$$y^* = \sum_{i=0}^{M} b_i 2^{-i} + \begin{cases} b_{M+1} 2^{-M} \text{ for roundoff,} \\ 0 \text{ for chopping .} \end{cases} \tag{36}$$

Note that b_{M+1} can be written as

$$b_{M+1} = I_{2^N} [2^{M+N} b^*] . \tag{37}$$

The value y associated with y^* can be written as

$$y = b + \epsilon \tag{38}$$

where ϵ is the value of the roundoff or chopping error. The relation between y and y^* is the same as that between b and b^* in (29)–(30). Thus, for non-negative numbers $y \geq 0$,

$$y^* = y \tag{39}$$

and for negative numbers $y < 0$,

$$y^* = 2 + y\,. \tag{40}$$

ROUNDOFF AND CHOPPING

The error can be calculated from (29)–(38) as

$$\begin{aligned} \epsilon &= y - b \\ &= y^* - b^* \\ &= -r^* 2^{-M} + \begin{cases} b_{M+1} 2^{-M} \text{ for roundoff,} \\ 0 \text{ for chopping,} \end{cases} \end{aligned} \tag{41}$$

where

$$r^* 2^{-M} = 2^{-M-N} [u]_{2^N} \tag{42}$$

and

$$b_{M+1} 2^{-M} = 2^{-M} I_{2^N}[u] \tag{43}$$

for

$$u = -2^{M+N}, \ldots, -1, 0, 1, \ldots, 2^{M+N}-1\,. \tag{44}$$

EFFECT OF MULTIPLICATION COEFFICIENT

Consider the product

$$b = xa \tag{45}$$

where

$$x = k2^{-K}\,, \tag{46}$$

$$a = \ell 2^{-L}\,, \tag{47}$$

and

$$K + L = M + N\,. \tag{48}$$

The integer k can take on the values

$$k = -2^K, \ldots, 2^K-2, 2^K-1 . \tag{49}$$

The intergers ℓ are similarly described. (Replace k by ℓ and K by L in (49).) Then, from (31)

$$u = k\ell . \tag{50}$$

Consider values of ℓ of the form

$$\ell = \ell' 2^\delta \tag{51}$$

were ℓ' is a constant odd number which can take on the values

$$\ell' = \pm 1, \pm 3, \ldots, \pm(2^{L-\delta}-1) \tag{52}$$

when $\delta = 0, 1, \ldots, L-1$. Note that ℓ' can take on the value $\ell' = -1$ when $\delta = L$. The values of ℓ of the form (51) thus span the set of all possible non-zero values of ℓ as defined by (47).

The values of the error ϵ in (41) can be evaluated through the substitution

$$u = k\ell' 2^\delta . \tag{53}$$

Simplification of the error equations can be made through introduction of the parameter ν where $\nu = N-\delta$. Thus, for example, for non-negative numbers u (or $k\ell' \geq 0$),

$$\begin{aligned} r^* 2^{-M} &= 2^{-M-N}[k\ell' 2^\delta]_{2N} \\ &= 2^{-M-N} 2^\delta [k\ell']_{2N-\delta} \\ &= 2^{-M-\nu}[k\ell']_{2^\nu} \end{aligned} \tag{54}$$

and

$$\begin{aligned} b_{M+1} 2^{-M} &= 2^{-M} I_{2N}[k\ell' 2^\delta] \\ &= 2^{-M} I_{2^\nu}[k\ell'] . \end{aligned} \tag{55}$$

The same results are obtained for negative numbers u (or $k\ell' < 0$).

These equations are valid when no overflow occurs and also for the specified values of ν and δ. Overflow can occur for roundoff of positive numbers ($k\ell > 0$). It cannot occur for roundoff of negative numbers ($k\ell < 0$) for TC. Overflow occurs when the unmodified product b takes the form

$$xa = 0.11\ldots11r_2r_3\ldots r_N \cdot \tag{56}$$

Thus, the values k that result in overflow are those that satisfy the inequality

$$k\ell \geq (2^{M+1}-1)2^{N-1} . \tag{57}$$

The effect of overflow is not mirrored in the equations however. The interesting values of ν are given by the inequalities $1 \leq \nu \leq N$ when $L \geq N$, and $(N-L) < \nu \leq N$ when $L < N$. Note that, when $\delta = L$, the coefficient value is equal to zero and no errors occur.

TC ERROR PROPERTIES

In the following sections, the integer form e of the error will be used. It is defined through (41)–(51) as the following:

$$e = \epsilon 2^{M+\nu}$$

$$= -[k\ell']_{2^\nu} + \begin{cases} 2^\nu I_{2^\nu}[k\ell'] \text{ for roundoff,} \\ 0 \text{ for chopping .} \end{cases} \tag{58}$$

The following error properties become evident.

PROPERTY 1: Consider values of k of the form $k = k' + p2^\nu$ where the integers k' and p are confined to the ranges

$$0 \leq k' < 2^\nu \tag{59}$$

and

$$-2^{K-\nu} \leq p < 2^{K-\nu} . \tag{60}$$

For constant coefficient a and associated factors ℓ' and ν, all values of k with the same constant k' map onto the same value of the error e. Furthermore, the mapping from k' to e is one-to-one with the mapping specified by (58).

In much of the following analysis, this property finds use whenever the resulting relation

$$e(k') = e(k' + p2^\nu) \tag{61}$$

is employed. It is easily seen that the range of error values e which can occur are given as

$$-2^{\nu-1}+1 \leq e \leq 2^{\nu-1} \tag{62}$$

for roundoff, and

$$-2^{\nu}+1 \leq e \leq 0 \tag{63}$$

for chopping.

PROPERTY 2: Consider the two coefficient values $a_+ > 0$ and $a_- < 0$ where $a_+ = -a_-$. Also let $e_\gamma(k)$ describe the mapping from k to e for a_γ constant where $\gamma = +$ or $-$. Then $e_+(k) = e_-(-k)$ for any integer $k \in (-\infty, \infty)$. In particular, $e_-(k') = e_+(2^\nu - k')$ for $k' = 1, 2, \ldots, 2^\nu - 1$, and $e_-(k') = e_+(k')$ for $k' = 0$.

For constant coefficient value, a sequence of error values results when the multiplier input is stepped sequentially through the values $k = \ldots, -1, 0, 1, \ldots$. This property states that the resulting sequence of error values for a_+ is the mirror image (reflected around $k = 0$) of the sequence that results for $a_- = -a_+$.

A graphical example of the mappings from the multiplier input to the error is shown in Fig. 3 for positive coefficient values and roundoff. The patterns are grouped in columns according to the parameter ν where $\nu = 1, 2, 3, 4$ as shown. Only one pattern results for $\nu = 1$, two for $\nu = 2$, four for $\nu = 3$, etc. Not all patterns for $\nu = 4$ are shown. Other features of the patterns that result are illustrated here. First, each possible error value is mapped onto by a set of multiplier input values which differ by 2^ν. Second, ν describes the class of coefficient values that result in the same number of possible error values (and the same error model for large standard deviation of the multiplier input as shown in [3]). Third, the mapping from the multiplier input to the error value is dependent on the value of the coefficient for a given ν. And fourth, for some word length combinations, the same mapping can occur for two different coefficient values for a given ν.

The patterns that occur for chopping are the same as for roundoff for each coefficient value. The differences lie in a shift both in the horizontal (multiplier input variable) and vertical (multiplication error) direction. Thus, the same comments hold regarding the ν class, number of error values, reflection and other properties.

IV. MULTIPLIER INPUT STATISTICS

In this section the analog signal quantization process is reviewed. Relevant quantizer output (hence filter input) statistics are presented ending with the Gaussian case. Statistics are the filter input mean and the input auto- and cross-correlation for zero and non-zero time lag. Use of the PSF is demonstrated and the notation applied to the rest of the paper is discussed.

UNIVARIATE CHARACTERISTICS

Let x be a discrete random variable (r.v.) which is obtained from a continuously distributed r.v. $\tilde{x}$ by quantization with roundoff. That is, let $x = kq$ where the integer k is chosen such that $q(k-\frac{1}{2}) \leq \tilde{x} < q(k+\frac{1}{2})$ and $q = 2^{-K}$ is the quantization step size. The A/D converter used has an output word length of K bits plus sign with TC number representation. This restricts the values of the A/D converter output to correspond to the range denoted by $k2^{-K}$ where $k = -2^K, -2^K+1, \ldots, 2^K-1$. In the following, it is mathematically convenient to assume x (and hence k) is not bounded. In order to practically realize this assumption, the A/D converter input can be made small enough in most cases so that the effect of quantizer overflow on the derived statistics is negligible.

In the following it is assumed that $\tilde{x}$ is statistically stationary and has a probability density function (p.d.f.) $f_{\tilde{x}}(\tilde{x})$ associated with it. This p.d.f. is also assumed to be continuous everywhere on the real line. The probabilities of occurrence of each k are written as $P_x(k)$ and are defined as

$$P_x(k) = \int_{q(k-\frac{1}{2})}^{q(k+\frac{1}{2})} f_{\tilde{x}}(\tilde{x})d\tilde{x} \tag{64}$$

for $k = 0, \pm1, \pm2, \ldots$. The dependence on K is assumed understood.

Associated with the p.d.f. $f_{\tilde{x}}(\tilde{x})$ is the characteristic function $Q_{\tilde{x}}(\omega)$ defined by the Fourier transform

$$Q_{\tilde{x}}(\omega) = \int_{-\infty}^{\infty} f_{\tilde{x}}(\xi)\exp(j\omega\xi)d\xi \ . \tag{65}$$

The special case that will be used is that of the Gaussian p.d.f. $f_{\tilde{x}}(\tilde{x})$ where

$$f_{\tilde{x}}(\tilde{x}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\tilde{x}-\mu)^2}{2\sigma^2}\right\} \tag{66}$$

with characteristic function given by

$$Q_{\tilde{x}}(\omega) = \exp\left\{-\frac{\sigma^2\omega^2}{2} + j\mu\omega\right\} . \tag{67}$$

The expression $P_x(k)$ is well-behaved for any real k. Its Fourier transform $Q_x(\omega)$ is given by

$$Q_x(\omega) = \int_{-\infty}^{\infty} P_x(\xi)\exp(j\omega\xi)d\xi \tag{68}$$

$$= \text{sinc}(\omega/2\pi)Q_{\tilde{x}}(\omega/q)$$

where $\text{sinc}(x) = \sin(\pi x)/\pi x$.

Note that [4]

$$\left\{\frac{d}{d\omega}Q_x(\omega)\right\}_{\omega=0} = j\mu/q \tag{69}$$

and

$$\left\{\frac{d^2}{d\omega^2}Q_x(\omega)\right\}_{\omega=0} = -\frac{1}{q^2}R_{\tilde{x}}(0) - \frac{1}{12} . \tag{70}$$

MEAN MULTIPLIER INPUT

The mean multiplier input μ_x can be obtained as follows:

$$\mu_x = E[x]$$

$$= 2^{-K}E[k] \tag{71}$$

$$= q\sum_{k=-\infty}^{\infty} kP_x(k) .$$

Since the function $kP_x(k)$ is well behaved for any real k, the PSF can be employed on (71) to yield

$$\mu_x = -jq \sum_{s=-\infty}^{\infty} \left\{ \frac{d}{d\omega} Q_x(\omega) \right\}_{\omega=2\pi s} \tag{72}$$

$$= \vec{\mu}_x + \downarrow\mu_x$$

where, from (69),

$$\vec{\mu}_x = \mu \tag{73}$$

and

$$\downarrow\mu_x = -jq \sum_{\substack{s=-\infty \\ s\neq 0}}^{\infty} \frac{(-1)^s}{2\pi s} Q_{\tilde{x}}(2\pi s/q) \ . \tag{74}$$

Note that, aside from stationarity and continuity assumptions, no special form of $f_{\tilde{x}}(\tilde{x})$ is assumed at this point. The Gaussian form of $\downarrow\mu_x$ is obtained through substitution of (67) into (74). The result is

$$\downarrow\mu_x = \frac{q}{\pi} \sum_{s=1}^{\infty} \frac{(-1)^s}{s} \exp\left\{-2(\pi s)^2(\sigma/q)^2\right\} \sin(2\pi s\mu/q) \ . \tag{75}$$

NOTATION AND ASSUMPTIONS

There are some features of $\vec{\mu}_x$ and $\downarrow\mu_x$ which are present in subsequent equations where the arrow notation is used. As can be seen from (75) the function $\downarrow\mu_x$ is dependent on two parameters, μ and σ, each in relation to the quantization step size q. The function can be made to approach arbitrarily close to zero through a suitable increase in the parameter σ for any constant value of μ.

Mathematical limits of μ_x can be taken in either one of two ways. The first way is to increase σ and keep μ constant. The second way is to increase σ and keep the ratio σ/μ constant. The first limit has the distinction that $\vec{\mu}_x$ is a constant. For the second limit $\vec{\mu}_x$ increases as σ. It will be assumed that the horizontal arrow denotes a variable which is either constant or depends in a well-behaved manner on μ and/or σ for a finite μ and σ. The vertical arrow denotes a variable which tends to zero as σ increases regardless of the assumptions about μ.

As with the mean quantizer output, subsequent cases will occur where a progression is made from a statistic expressed as an expected value to the asymptotic ($\rightarrow$) and decreasing ($\downarrow$) terms. It will be assumed without comment that the functions involved are well-behaved and that the PSF was used to arrive at the final results.

MULTIPLIER INPUT SECOND MOMENT

The second moment of the multiplier input is $R_x(0)$. It is the autocorrelation function for zero time lag. It can be obtained as follows:

$$\begin{aligned} R_x(0) &= E[x^2] \\ &= 2^{-2K} E[k^2] \\ &= q^2 \sum_{k=-\infty}^{\infty} k^2 P_x(k) \\ &= -q^2 \sum_{s=-\infty}^{\infty} \left\{ \frac{d^2}{d\omega^2} Q_x(\omega) \right\}_{\omega=2\pi s} \\ &= \vec{R}_x(0) + \downarrow R_x(0) \end{aligned} \tag{76}$$

where, from (70)

$$\begin{aligned} \vec{R}_x(0) &= R_{\tilde{x}}(0) + q^2/12 \\ &= \sigma^2 + \mu^2 + q^2/12 , \end{aligned} \tag{77}$$

and, for the Gaussian case

$$\downarrow R_x(0) = 2q^2 \sum_{s=1}^{\infty} (-1)^s \left[\left\{ \frac{1}{2(\pi s)^2} + 2(\sigma/q)^2 \right\} \cos(2\pi s\mu/q) + \left\{ \frac{1}{\pi s} (\mu/q) \right\} \sin(2\pi s(\mu/q)) \right] \exp(-2(\pi s)^2(\sigma/q)^2) . \tag{78}$$

BIVARIATE CHARACTERISTICS

By analogy with (64), the joint probability of occurrence of a pair of multiplier inputs (k_1,k_2) is written as $P_{x_1x_2}(k_1,k_2)$ which can be specified in the same way as the set of probabilities $\{P_x(k)\}$. Let the multiplier input r.v.'s x_1 and x_2 be obtained from the joint r.v.'s $\tilde{x}_1$ and $\tilde{x}_2$ by quantization with roundoff. Assume $\tilde{x}_1$ and $\tilde{x}_2$ are stationary and

have a joint p.d.f. $f_{\tilde{x}_1\tilde{x}_2}(\tilde{x}_1,\tilde{x}_2)$ that is continuous everywhere in the interval $-\infty < \tilde{x}_1,\tilde{x}_2 < \infty$. Further assume $\tilde{x}_1$ and $\tilde{x}_2$ are correlated with normalized correlation coefficient $\rho = C_{\tilde{x}_1\tilde{x}_2}(\cdot)/(\sigma_1\sigma_2)$ where $|\rho| < 1$ and $\sigma_1 = \sigma_{\tilde{x}_1}$ and $\sigma_2 = \sigma_{\tilde{x}_2}$. Thus

$$P_{x_1x_2}(k_1,k_2) = \int_{q(k_1-\frac{1}{2})}^{q(k_1+\frac{1}{2})} \int_{q(k_2-\frac{1}{2})}^{q(k_2+\frac{1}{2})} f_{\tilde{x}_1\tilde{x}_2}(\tilde{x}_1,\tilde{x}_2)d\tilde{x}_1d\tilde{x}_2 \tag{79}$$

for $k_1,k_2 = 0, \pm1, \pm2, \ldots$.

Associated with the p.d.f. $f_{\tilde{x}_1\tilde{x}_2}(\tilde{x}_1,\tilde{x}_2)$ is the characteristic function $Q_{\tilde{x}_1\tilde{x}_2}(\omega_1,\omega_2)$ defined by the Fourier transform

$$Q_{\tilde{x}_1\tilde{x}_2}(\omega_1,\omega_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{\tilde{x}_1\tilde{x}_2}(\xi_1,\xi_2)\exp(j\omega_1\xi_1 + j\omega_2\xi_2)d\xi_1d\xi_2 \,. \tag{80}$$

The Gaussian case will be used here also where $\sigma_{\tilde{x}_1} = \sigma_{\tilde{x}_2} = \sigma$ and $\mu_{\tilde{x}_1} = \mu_{\tilde{x}_2} = \mu$. Thus

$$f_{\tilde{x}_1\tilde{x}_2}(\tilde{x}_1,\tilde{x}_2) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \text{ times} \tag{81}$$

$$\exp\left\{-\frac{1}{2\sigma^2(1-\rho^2)}\left[(\tilde{x}_1-\mu)^2 - 2\rho(\tilde{x}_1-\mu)(\tilde{x}_2-\mu) + (\tilde{x}_2-\mu)^2\right]\right\}$$

with characteristic function given by

$$Q_{\tilde{x}_1\tilde{x}_2}(\omega_1,\omega_2) = \exp\left\{-\frac{\sigma^2}{2}(\omega_1^2 + 2\rho\omega_1\omega_2 + \omega_2^2) + j\mu(\omega_1+\omega_2)\right\} \,. \tag{82}$$

The function $P_{x_1x_2}(k_1,k_2)$ is well-behaved for any pair of real values (k_1,k_2). Its Fourier transform $Q_{x_1x_2}(\omega_1,\omega_2)$ is given by

$$\begin{aligned} Q_{x_1x_2}(\omega_1,\omega_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P_{x_1x_2}(\xi_1,\xi_2)\exp(j\omega_1\xi_1 + j\omega_2\xi_2)d\xi_1d\xi_2 \\ &= \text{sinc}(\omega_1/2\pi)\text{sinc}(\omega_2/2\pi)Q_{\tilde{x}_1\tilde{x}_2}(\omega_1/q,\omega_2/q) \,. \end{aligned} \tag{83}$$

Note that [4]

$$\left\{\frac{d}{d\omega_1} Q_{x_1x_2}(\omega_1,\omega_2)\right\}_{\omega_1,\omega_2=0} = j\mu/q \tag{84}$$

and

$$\left\{\frac{d^2}{d\omega_1 d\omega_2} Q_{x_1x_2}(\omega_1,\omega_2)\right\}_{\omega_1,\omega_2=0} = -\frac{1}{q^2} R_{\tilde{x}_1\tilde{x}_2}(0) \ . \tag{85}$$

MULTIPLIER INPUT AUTOCORRELATION/CROSS-CORRELATION

The multiplier input autocorrelation function for non-zero time lag $\{\eta T; \eta \neq 0\}$ is $R_x(\eta T)$. It can be written equivalently as $R_{x_1x_2}(0)$ which is the cross-correlation function for two quantizers with equal quantile step size q and zero time lag. The function presented here is valid only when $|\rho| < 1$. If $\rho = 1$, the autocorrelation function is equal to the second moment (76). Thus,

$$\begin{aligned} R_{x_1x_2}(0) &= E[x_1x_2] \\ &= 2^{-2K} E[k_1k_2] \\ &= q^2 \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} k_1k_2 P_{x_1x_2}(k_1,k_2) \\ &= -q^2 \sum_{s_1=-\infty}^{\infty} \sum_{s_2=-\infty}^{\infty} \left\{\frac{d^2}{d\omega_1 d\omega_2} Q_{x_1x_2}(\omega_1,\omega_2)\right\}_{\substack{\omega_1=2\pi s_1 \\ \omega_2=2\pi s_2}} \\ &= \vec{R}_{x_1x_2}(0) + \downarrow R_{x_1x_2}(0) \end{aligned} \tag{86}$$

where, from (85)

$$\begin{aligned} \vec{R}_{x_1x_2}(0) &= R_{\tilde{x}_1\tilde{x}_2}(0) \\ &= \rho\sigma^2 + \mu^2 \end{aligned} \tag{87}$$

and, for the Gaussian case

$$\downarrow R_{x_1x_2}(0) = \frac{q^2}{2\pi^2} \sum_{s_1=1}^{\infty} \sum_{s_2=1}^{\infty} \frac{(-1)^{s_1+s_2}}{s_1 s_2} \exp\left\{-2(\sigma/q)^2\pi^2(s_1^2 - 2\rho s_1 s_2 + s_2^2)\right\} \cos\left\{2\pi(s_1 - s_2)\mu/q\right\}$$

$$- \frac{q^2}{2\pi^2} \sum_{s_1=1}^{\infty} \sum_{s_2=1}^{\infty} \frac{(-1)^{s_1+s_2}}{s_1 s_2} \exp\left\{-2(\sigma/q)^2\pi^2(s_1^2 + 2\rho s_1 s_2 + s_2^2)\right\} \cos\left\{2\pi(s_1+s_2)\mu/q\right\}$$

$$+ 4\rho\sigma^2 \sum_{s=1}^{\infty} (-1)^s \exp\left\{-2(\sigma/q)^2(\pi s)^2\right\} \cos\left\{2\pi s\mu/q\right\} \qquad (88)$$

$$+ \frac{2}{\pi}\mu q \sum_{s=1}^{\infty} \frac{(-1)^s}{s} \exp\left\{-2(\sigma/q)^2(\pi s)^2\right\} \sin\left\{2\pi s\mu/q\right\} .$$

V. ERROR STATISTICS

The equations of the error statistics are derived in this section. The asymptotic and decreasing terms are separated and identified. The asymptotic terms are found to be independent of the p.d.f. shape. The decreasing terms are left general in that they contain the forms Q_x and Q_{xx}. This section is intended to show the derivations and results with a minimum of explanation or discussion. Gaussian forms of the decreasing terms are presented in the next section. Properties of the asymptotic terms are then discussed in the subsequent section.

If the reader is concerned primarily with using the equations for Gaussian quantizer inputs, he should use only the asymptotic terms from this section and the decreasing forms from the next section.

PROBABILITIES

Consider the case of a single multiplier with multiplication coefficient a, associated parameter ν, and word lengths K, L, M, N fixed with K+L = M+N. By Property 1, the values of k which map onto a particular value of the error e are given by the equation $k = k'+p2^{\nu}$ where the relation between k' and e is given by (58). In line with the relaxation of restrictions in the previous section, it will be assumed k (and hence p) is not bounded. The probability of occurrence of each e, $P_\epsilon(e) = P_\epsilon(e; a, \nu)$, can be written in terms of the probability of occurrence of each k, $P_x(k)$ which maps onto it. The factors a, ν and the word lengths K, L, M, and N all condition the values computed of $P_\epsilon(e)$ for any given p.d.f. $f_{\tilde{x}}(\tilde{x})$. This dependence is assumed understood in the following. Thus,

$$P_\epsilon(e) = \sum_{p=-\infty}^{\infty} P_x(k'+p2^{\nu}) \tag{89}$$

$$= 2^{-\nu} \sum_{s=-\infty}^{\infty} Q_x(\omega_s) \exp(-jk'\omega_s)$$

$$= \vec{P}_\epsilon + \downarrow P_\epsilon(e)$$

where

$$\vec{P}_\epsilon = 2^{-\nu}\,, \tag{90}$$

$$\downarrow P_\epsilon(e) = 2^{-\nu} \sum_{\substack{s=-\infty \\ s\neq 0}}^{\infty} Q_x(\omega_s) \exp(-jk'\omega_s)\,, \tag{91}$$

and

$$\omega_s = 2\pi s/2^{\nu} . \tag{92}$$

The probabilities can also be derived for joint r.v.'s. Let a_1 and a_2 be the constant coefficients of two separate multipliers with K, L, M and N the same for both. This results in two values of ν, ν_1 and ν_2, which are not necessarily equal.

Consider the sets of integers $k_1 = k'_1 + p_1 2^{\nu_1}$ and $k_2 = k'_2 + p_2 2^{\nu_2}$ where $k'_1, k'_2, \nu_1, \nu_2, p_1$ and p_2 are integers with properties as discussed above. These values of k_1 and k_2 map onto a pair of integers e_1 and e_2 with values determined by k'_1 and k'_2 respectively by (58). The probability of occurrence of each pair of errors (e_1, e_2) is written as $P_{\epsilon_1\epsilon_2}(e_1,e_2) = P_{\epsilon_1\epsilon_2}(e_1,e_2; a_1,a_2,\nu_1,\nu_2)$. This probability can be written in terms of the probability of occurrence $P_{x_1x_2}(k_1,k_2)$ of each pair (k_1,k_2) which maps onto it. Thus

$$P_{\epsilon_1\epsilon_2}(e_1,e_2) = \sum_{p_1=-\infty}^{\infty} \sum_{p_2=-\infty}^{\infty} P_{x_1x_2}(k'_1+p_1 2^{\nu_1}, k'_2+p_2 2^{\nu_2})$$

$$= 2^{-\nu_1-\nu_2} \sum_{s_1=-\infty}^{\infty} \sum_{s_2=-\infty}^{\infty} Q_{x_1x_2}(\omega_{s_1},\omega_{s_2}) \text{ times}$$

$$\exp(-jk'_1\omega_{s_1} - jk'_2\omega_{s_2}) \tag{93}$$

$$= \vec{P}_{\epsilon_1\epsilon_2} + \downarrow P_{\epsilon_1\epsilon_2}(e_1,e_2)$$

where

$$\vec{P}_{\epsilon_1\epsilon_2} = 2^{-\nu_1-\nu_2} , \tag{94}$$

$$\downarrow P_{\epsilon_1\epsilon_2}(e_1,e_2) = 2^{-\nu_1-\nu_2} \sum_{\substack{s_1=-\infty \\ (s_1,s_2)\neq(0,0)}}^{\infty} \sum_{s_2=-\infty}^{\infty} Q_{x_1x_2}(\omega_{s_1},\omega_{s_2}) \text{ times}$$

$$\exp(-jk'_1\omega_{s_1} - jk'_2\omega_{s_2}) \tag{95}$$

and

$$\omega_{s_i} = 2\pi s_i/2^{\nu_i} \text{ for } i = 1,2 . \tag{96}$$

MEAN ERROR

Through (89) the mean error μ_ϵ can be expressed as

$$\begin{aligned}\mu_\epsilon &= E[\epsilon] \\ &= 2^{-M-\nu} E[e] \\ &= 2^{-M-\nu} \sum_e e\, P_\epsilon(e) \\ &= \vec{\mu}_\epsilon + \downarrow\mu_\epsilon\end{aligned} \tag{97}$$

where

$$\vec{\mu}_\epsilon = 2^{-M-2\nu} \sum_e e$$

$$= \begin{cases} 2^{-M-\nu-1} \text{ for TCR (Two's-Complement Roundoff)} \\ 2^{-M-1}(2^{-\nu}-1) \text{ for TCC (Two's-Complement Chopping)} \end{cases} \tag{98}$$

and

$$\downarrow\mu_\epsilon = 2^{-M-\nu} \sum_e e\, \downarrow P_\epsilon(e) \ . \tag{99}$$

The sum (98) is evaluated through use of (62) and (63). For convenience the limits on e are not shown because they depend on whether roundoff or chopping is used.

The mapping from k′ to e is one-to-one. Hence, the error can be written as a function of k′; namely e = e(k′). Also, a summation over all possible values of e is equivalent to a summation over all possible values of k′. Thus, through (91) the term $\downarrow\mu_\epsilon$ can be written as

$$\begin{aligned}\downarrow\mu_\epsilon &= 2^{-M-\nu} \sum_e e(k') \sum_{\substack{s=-\infty \\ s\neq 0}}^{\infty} 2^{-\nu} Q_X(\omega_s) \exp(-jk'\omega_s) \\ &= 2^{-M-2\nu} \sum_{\substack{s=-\infty \\ s\neq 0}}^{\infty} Q_X(\omega_s) D_s\end{aligned} \tag{100}$$

(contd)

$$= 2^{-M-2\nu+1} \sum_{s=1}^{\infty} \mathrm{Re}\left\{ Q_x(\omega_s) D_s \right\}$$

where

$$D_s = \sum_{k'=0}^{2^\nu - 1} e(k') \exp(-jk'\omega_s) \ . \tag{101}$$

The quantity D_s is evaluated in the Appendix and is given as

$$D_s = \begin{cases} 2^{\nu-1} & \text{for } s \equiv 0 \bmod 2^\nu \\ (-1)^s 2^{\nu-1} \dfrac{\exp((\operatorname{sgn} a) j\omega_s \lambda) - 1}{\cos(\omega_s \lambda) - 1} & \text{otherwise} \end{cases} \tag{102}$$

for TCR, and

$$D_s = \begin{cases} 2^{\nu-1}(1-2^\nu) & \text{for } s \equiv 0 \bmod 2^\nu \\ 2^{\nu-1} \dfrac{\exp((\operatorname{sgn} a) j\omega_s \lambda) - 1}{\cos(\omega_s \lambda) - 1} & \text{otherwise} \end{cases} \tag{103}$$

for TCC. Note that D_s can be a complex quantity and that it is a function of the parameter λ which satisfies the relation

$$1 \equiv |\ell'| \lambda \bmod 2^\nu \tag{104}$$

where ℓ' is in turn related to the value of the coefficient a through (47)–(52).

$R_\epsilon(0)$

The second moment of the multiplication error for one multiplier is denoted by $R_\epsilon(0)$. Thus

$$R_\epsilon(0) = E[\epsilon^2]$$

$$= 2^{-2M-2\nu} E[e^2]$$

$$= 2^{-2M-2\nu} \sum_{e} e^2 P_\epsilon(e) \tag{105}$$

(contd)

$$= \vec{R}_\epsilon(0) + \downarrow R_\epsilon(0)$$

where

$$\vec{R}_\epsilon(0) = 2^{-2M-3\nu} \sum_e e^2$$

$$= \begin{cases} \frac{2^{-2M}}{6}\left[\frac{1}{2} + 2^{-2\nu}\right] & \text{for TCR} \\ \frac{2^{-2M}}{6}\{2 - 3\cdot 2^{-\nu} + 2^{-2\nu}\} & \text{for TCC} \end{cases} \tag{106}$$

and

$$\downarrow R_\epsilon(0) = 2^{-2M-2\nu} \sum_e e^2 \downarrow P_\epsilon(e)$$

$$= 2^{-2M-2\nu} \sum_e e^2(k') \sum_{\substack{s=-\infty \\ s\neq 0}}^{\infty} 2^{-\nu} Q_X(\omega_s) \exp(-jk'\omega_s)$$

$$= 2^{-2M-3\nu+1} \sum_{s=1}^{\infty} \mathrm{Re}\left\{Q_X(\omega_s)F_s\right\}. \tag{107}$$

The quantity F_s defined by

$$F_s = \sum_e e^2(k') \exp(-jk'\omega_s) \tag{108}$$

is evaluated in the Appendix and is given by

$$F_s = \begin{cases} \frac{1}{12}(2^{3\nu} + 2\cdot 2^{\nu}) & \text{for} \quad s \equiv 0 \bmod 2^{\nu} \\ (-1)^s \frac{2^{\nu}}{1-\cos(\omega_s\lambda)} & \text{otherwise,} \end{cases} \tag{109}$$

for TCR, and

$$F_s = \begin{cases} \frac{1}{6}(2\cdot 2^{3\nu} - 3\cdot 2^{2\nu} + 2^{\nu}) & \text{for} \quad s \equiv 0 \bmod 2^{\nu} \\ \\ \frac{2^{2\nu-1}[\exp((\text{sgn } a)j\omega_s\lambda)-1]+2^{\nu}}{1-\cos(\omega_s\lambda)} & \text{otherwise} \end{cases} \tag{110}$$

for TCC. It can be complex and is dependent on the coefficient value through the parameter λ defined by (104) which is the same relation as used for D_s.

$R_{x\epsilon}$ for $\rho = 1.0$

The joint moment of a r.v. x and the corresponding error ϵ for the same multiplier is denoted by $R_{x\epsilon}$ where

$$R_{x\epsilon} = E[x\epsilon]$$

$$= q2^{-M-\nu} E[ke] \tag{111}$$

and

$$E[ke] = \sum_{k=-\infty}^{\infty} k\, e(k)\, P_x(k)$$

$$= \sum_{p=-\infty}^{\infty} \sum_{k'=0}^{2^{\nu}-1} (k'+p2^{\nu})\, e(k'+p2^{\nu})\, P_x(k'+p2^{\nu})$$

$$= \sum_{k'=0}^{2^{\nu}-1} e(k') \sum_{p=-\infty}^{\infty} (k'+p2^{\nu})\, P_x(k'+p2^{\nu}) \tag{112}$$

$$= \sum_{k'=0}^{2^{\nu}-1} e(k') \sum_{s=-\infty}^{\infty} 2^{-\nu}(-j)\left\{\frac{d}{d\omega}Q_x(\omega)\right\}_{\omega=\omega_s} \exp(-jk'\omega_s)$$

$$= -j2^{-\nu} \sum_{s=-\infty}^{\infty} \left\{\frac{d}{d\omega}Q_x(\omega)\right\}_{\omega=\omega_s} D_s$$

$$= 2^{-\nu}\,\mu D_0/q - j2^{-\nu} \sum_{\substack{s=-\infty \\ s\neq 0}}^{\infty} \left\{\frac{d}{d\omega}Q_x(\omega)\right\}_{\omega=\omega_s} D_s \,.$$

The third step in (112) is possible since $e(k') = e(k'+p2^{\nu})$ and since the mapping from k' to e is one-to-one. The last step is obtained through use of (69). Thus,

$$R_{x\epsilon} = \vec{R}_{x\epsilon} + \downarrow R_{x\epsilon} \tag{113}$$

where

$$\vec{R}_{x\epsilon} = \mu 2^{-M-2\nu} D_0$$

$$= \begin{cases} \mu 2^{-M-\nu-1} & \text{for TCR} \\ \mu 2^{-M-\nu-1}(1-2^{\nu}) & \text{for TCC,} \end{cases} \tag{114}$$

and

$$\downarrow R_{x\epsilon} = -jq2^{-M-2\nu} \sum_{\substack{s=-\infty \\ s\neq 0}}^{\infty} \left\{ \frac{d}{d\omega} Q_x(\omega) \right\}_{\omega=\omega_s} D_s \,. \tag{115}$$

Note that this is the same joint moment as for the r.v. x_1 associated with multiplier a_1 and the error ϵ_2 associated with multiplier a_2 when $\rho = 1.0$. In this case it is necessary to make the assignments $\nu = \nu_2$ and $s = s_2$ so that $\omega_s = 2\pi s_2/2^{\nu 2}$.

$R_{x_1\epsilon_2}$ for $|\rho| < 1.0$

Let x_1 be the input for the multiplier with coefficient a_1 and ϵ_2 be the error for the multiplier with coefficient a_2. Furthermore, let $|\rho| < 1$. The joint moment of the r.v.'s x_1 and ϵ_2 is denoted by $R_{x_1\epsilon_2}$ where

$$R_{x_1\epsilon_2} = E[x_1\epsilon_2]$$

$$= q2^{-M-\nu_2} E[k_1 e_2] \tag{116}$$

and

$$E[k_1 e_2] = \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} k_1 e_2(k_2)\, P_{x_1 x_2}(k_1,k_2)$$

$$= \sum_{k_1=-\infty}^{\infty} \sum_{p_2=-\infty}^{\infty} \sum_{k_2'=0}^{2^{\nu_2}-1} k_1\, e_2(k_2'+p_2 2^{\nu_2})\, P_{x_1 x_2}(k_1,k_2'+p_2 2^{\nu_2})$$

$$= \sum_{k_2'=0}^{2^{\nu_2}-1} e_2(k_2') \sum_{k_1=-\infty}^{\infty} \sum_{p_2=-\infty}^{\infty} k_1\, P_{x_1 x_2}(k_1,k_2'+p_2 2^{\nu_2}) \qquad (117)$$

$$= \sum_{k_2'=0}^{2^{\nu_2}-1} e_2(k_2') 2^{-\nu_2} \quad \text{times}$$

$$\sum_{s_1=-\infty}^{\infty} \sum_{s_2=-\infty}^{\infty} (-j) \left\{ \frac{d}{d\omega_1} Q_{x_1 x_2}(\omega_1,\omega_2) \right\}_{\substack{\omega_1=2\pi s_1 \\ \omega_2=\omega_{s_2}}} \exp(-jk_2'\omega_{s_2})$$

$$= -j2^{-\nu_2} \sum_{s_1=-\infty}^{\infty} \sum_{s_2=-\infty}^{\infty} \left\{ \frac{d}{d\omega_1} Q_{x_1 x_2}(\omega_1,\omega_2) \right\}_{\substack{\omega_1=2\pi s_1 \\ \omega_2=\omega_{s_2}}} D_{s_2}$$

$$= 2^{-\nu_2} \frac{\mu}{q} \{ D_{s_2} \}_{s_2=0}$$

$$-j2^{-\nu_2} \sum_{\substack{s_1=-\infty \\ (s_1,s_2)\neq(0,0)}}^{\infty} \sum_{s_2=-\infty}^{\infty} \left\{ \frac{d}{d\omega_1} Q_{x_1 x_2}(\omega_1,\omega_2) \right\}_{\substack{\omega_1=2\pi s_1 \\ \omega_2=\omega_{s_2}}} D_{s_2}$$

where

$$D_{s_2} = \sum_{k'_2=0}^{2^{\nu_2}-1} e_2(k'_2) \exp(-jk'_2\omega_{s_2}) \ . \tag{118}$$

Thus

$$R_{x_1\epsilon_2} = \vec{R}_{x_1\epsilon_2} + \downarrow R_{x_1\epsilon_2} \tag{119}$$

where

$$\vec{R}_{x_1\epsilon_2} = \mu 2^{-M-2\nu_2}\left\{D_{s_2}\right\} s_2=0 \tag{120}$$

$$= \begin{cases} \mu 2^{-M-\nu_2-1} & \text{for TCR} \\ \mu 2^{-M-\nu_2-1}(1-2^{\nu_2}) & \text{for TCC,} \end{cases} \tag{121}$$

and

$$\downarrow R_{x_1\epsilon_2} = -jq2^{-M-2\nu_2} \sum_{\substack{s_1=-\infty \\ (s_1,s_2)\neq(0,0)}}^{\infty} \sum_{s_2=-\infty}^{\infty} \left\{ \frac{d}{d\omega_1} Q_{x_1x_2}(\omega_1,\omega_2) \right\}_{\substack{\omega_1=2\pi s_1 \\ \omega_2=\omega_{s_2}}} D_{s_2} \ . \tag{122}$$

The joint moment of the r.v. x and the error ϵ for the same multiplier when $|\rho| < 1.0$ is obtained by simply letting $a_1 = a_2$ in the above.

$R_{\epsilon_1\epsilon_2}$ for $\rho = 1.0$

Let ϵ_1 and ϵ_2 be the multiplication errors for the multipliers with coefficients a_1 and a_2, respectively. Also, let $\rho = 1.0$ and $\nu_1 \geq \nu_2$. The joint moment of the r.v.'s ϵ_1 and ϵ_2 is denoted by $R_{\epsilon_1\epsilon_2}$ where

$$R_{\epsilon_1\epsilon_2} = E[\epsilon_1\epsilon_2]$$

$$= 2^{-2M-\nu_1-\nu_2} E[e_1e_2] \ . \tag{123}$$

The number of distinct states of the pair (e_1,e_2) is determined, in this case, by ν_1, since $\nu_1 \geq \nu_2$. The probability of occurrence of each state is the probability of occurrence of e_1 alone. Also, since $\rho = 1.0$, the error value e_2 is uniquely determined by the error value e_1. Thus,

$$R_{\epsilon_1\epsilon_2} = 2^{-2M-\nu_1-\nu_2} \sum_{e_1} e_1\, e_2(k_1'(e_1))\, P_{\epsilon_1}(e_1)$$

$$= \vec{R}_{\epsilon_1\epsilon_2} + \downarrow R_{\epsilon_1\epsilon_2} \tag{124}$$

where, from (89)

$$\vec{R}_{\epsilon_1\epsilon_2} = 2^{-2M-2\nu_1-\nu_2} \sum_{e_1} e_1\, e_2(k_1'(e_1))$$

$$= 2^{-2M-2\nu_1-\nu_2} \sum_{k_1'=0}^{2^{\nu_1}-1} e_1(k_1')\, e_2(k_1') \tag{125}$$

and

$$\downarrow R_{\epsilon_1\epsilon_2} = 2^{-2M-2\nu_1-\nu_2+1} \sum_{e_1} e_1\, e_2(k_1'(e_1)) \text{ times}$$

$$\sum_{s_1=1}^{\infty} \mathrm{Re}\left\{Q_x(\omega_{s_1}) \exp(-jk_1'(e_1)\omega_{s_1})\right\}$$

$$= 2^{-2M-2\nu_1-\nu_2+1} \sum_{k_1'=0}^{2^{\nu_1}-1} e_1(k_1')e_2(k_1') \quad \text{times} \tag{126}$$

$$\sum_{s_1=1}^{\infty} \mathrm{Re}\left\{Q_x(\omega_{s_1}) \exp(-jk_1'\omega_{s_1})\right\} .$$

Example plots of e_1 versus e_2 appear in [6]. Unfortunately, attempts at arriving at more suitable closed forms for $\vec{R}_{\epsilon_1\epsilon_2}$ and $\downarrow R_{\epsilon_1\epsilon_2}$ have been unsuccessful except for the special case $\epsilon_1 = \epsilon_2$ already derived. (For $\epsilon_1 = \epsilon_2$, see the equation for $R_\epsilon(0)$.) This is the reason for reverting to sums which are more easily computed using the index k_1' instead of e_1.

$R_{\epsilon_1\epsilon_2}$ for $|\rho| < 1.0$

Let ϵ_1 and ϵ_2 be the multiplication errors for the multipliers with coefficient values a_1 and a_2, respectively. Also, let $|\rho| < 1$ as for $R_{x_1x_2}$. The joint moment of the r.v.'s ϵ_1 and ϵ_2 is $R_{\epsilon_1\epsilon_2}$ where

$$\begin{aligned} R_{\epsilon_1\epsilon_2} &= E[\epsilon_1\epsilon_2] \\ &= 2^{-2M-\nu_1-\nu_2} E[e_1e_2] \\ &= 2^{-2M-\nu_1-\nu_2} \sum_{e_1} \sum_{e_2} e_1e_2 P_{\epsilon_1\epsilon_2}(e_1,e_2) \\ &= \vec{R}_{\epsilon_1\epsilon_2} + \downarrow R_{\epsilon_1\epsilon_2} \end{aligned} \tag{127}$$

where, from (93)

$$\vec{R}_{\epsilon_1\epsilon_2} = 2^{-2M-2\nu_1-2\nu_2} \sum_{e_1} \sum_{e_2} e_1e_2$$

$$= \begin{cases} 2^{-2M-\nu_1-\nu_2-2} & \text{for TCR} \\ 2^{-2M-\nu_1-\nu_2-2}(1-2^{\nu_1})(1-2^{\nu_2}) & \text{for TCC} \end{cases} \tag{128}$$

and

$$\downarrow R_{\epsilon_1\epsilon_2} = 2^{-2M-\nu_1-\nu_2} \sum_{e_1} \sum_{e_2} e_1 e_2 \downarrow P_{\epsilon_1\epsilon_2}(e_1,e_2) . \tag{129}$$

Through (95) the term $\downarrow R_{\epsilon_1\epsilon_2}$ can be written as

$$\downarrow R_{\epsilon_1\epsilon_2} = 2^{-2M-\nu_1-\nu_2} \sum_{e_1} \sum_{e_2} e_1(k'_1)\, e_2(k'_2) \text{ times}$$

$$\sum_{\substack{s_1=-\infty \\ (s_1,s_2)\neq(0,0)}}^{\infty} \sum_{s_2=-\infty}^{\infty} 2^{-\nu_1-\nu_2} Q_{x_1x_2}(\omega_{s_1},\omega_{s_2}) \exp(-jk'_1\omega_{s_1} - jk'_2\omega_{s_2}) \tag{130}$$

$$= 2^{-2(M+\nu_1+\nu_2)} \sum_{\substack{s_1=-\infty \\ (s_1,s_2)\neq(0,0)}}^{\infty} \sum_{s_2=-\infty}^{\infty} Q_{x_1x_2}(\omega_{s_1},\omega_{s_2})\, D_{s_1} D_{s_2} \ .$$

The autocorrelation of the error for one multiplier can be obtained from $R_\epsilon(0)$ for $\rho = 1.0$ and (130) for $|\rho| < 1.0$ by letting $a_1 = a_2$.

VARIANCES AND COVARIANCES

The variance of the multiplication error $C_\epsilon(0)$ is defined as

$$C_\epsilon(0) = R_\epsilon(0) - \mu_\epsilon^2$$

$$= \vec{C}_\epsilon(0) + \downarrow C_\epsilon(0) \tag{131}$$

where, from (97) and (105),

$$\vec{C}_\epsilon(0) = \vec{R}_\epsilon(0) - \vec{\mu}_\epsilon^2 \tag{132}$$

$$= \frac{2^{-2M}}{12}\left\{1 - 2^{-2\nu}\right\}$$

for TCR and TCC, and

$$\downarrow C_\epsilon(0) = \downarrow R_\epsilon(0) - 2\vec{\mu}_\epsilon \downarrow \mu_\epsilon - \downarrow \mu_\epsilon^2 \ . \tag{133}$$

The function $C_{x_1\epsilon_2}$ is the covariance of the multiplier input x_1 (for the multiplier with coefficient a_1) with the error ϵ_2 (for the multiplier with coefficient a_2). Similarly, the function $C_{x\epsilon}$ is the covariance of a multiplier input with the corresponding error. Both can be defined as the following (dropping the subscripts in the first case for brevity):

$$C_{x\epsilon} = R_{x\epsilon} - \mu_x\mu_\epsilon$$

$$= \vec{C}_{x\epsilon} + \downarrow C_{x\epsilon} \tag{134}$$

where, from (72), (97), (113) and (119),

$$\vec{C}_{x\epsilon} = \vec{R}_{x\epsilon} - \vec{\mu}_x\vec{\mu}_\epsilon \tag{135}$$

$$= 0$$

for TCR and TCC, and

$$\downarrow C_{x\epsilon} = \downarrow R_{x\epsilon} - \vec{\mu}_x\downarrow\mu_\epsilon - \downarrow\mu_x\vec{\mu}_\epsilon - \downarrow\mu_x\downarrow\mu_\epsilon \tag{136}$$

The covariance $C_{\epsilon_1\epsilon_2}$ of the multiplication error ϵ_1 (for the multiplier with coefficient a_1) with the multiplication error ϵ_2 (for the multiplier with coefficient a_2) and $|\rho| < 1.0$ is defined as

$$C_{\epsilon_1\epsilon_2} = R_{\epsilon_1\epsilon_2} - \mu_{\epsilon_1}\mu_{\epsilon_2} \tag{137}$$

$$= \vec{C}_{\epsilon_1\epsilon_2} + \downarrow C_{\epsilon_1\epsilon_2}$$

where, from (97) and (127)

$$\vec{C}_{\epsilon_1\epsilon_2} = \vec{R}_{\epsilon_1\epsilon_2} - \vec{\mu}_{\epsilon_1}\vec{\mu}_{\epsilon_2} \tag{138}$$

$$= 0$$

for TCR and TCC, and

$$\downarrow C_{\epsilon_1\epsilon_2} = \downarrow R_{\epsilon_1\epsilon_2} - \vec{\mu}_{\epsilon_1}\downarrow\mu_{\epsilon_2} - \vec{\mu}_{\epsilon_2}\downarrow\mu_{\epsilon_1} - \downarrow\mu_{\epsilon_1}\mu_{\epsilon_2} \tag{139}$$

When $\rho = 1.0$, the asymptotic portion $\vec{C}_{\epsilon_1\epsilon_2}$ is generally not equal to zero. However, the decreasing portion $\downarrow C_{\epsilon_1\epsilon_2}$ has the same form as (139).

The following comments regard the asymptotic correlation coefficient (ACC) $\vec{\rho}_{\epsilon_1\epsilon_2}$ defined for $\rho = 1.0$ as

$$\vec{\rho}_{\epsilon_1\epsilon_2} = \vec{C}_{\epsilon_1\epsilon_2} \Big/ \left\{ C_{\epsilon_1}(0)\, C_{\epsilon_2}(0) \right\}^{1/2} \quad . \tag{140}$$

(A similar form of the ACC was examined by Girard [6] and Parker and Girard [7]. They assumed a zero-mean error which is not the case for TCR.) First consider the case when $\nu_1 = \nu_2 = \nu$. For a given ν the ACC values vary from a maximum of 1.0 to values which can become quite small but which are non-zero. Table 1 shows the computed ACC values which result when $\nu = 5$ and for two coefficient values ℓ'_1. Note that the same ACC values result for both cases but they are permuted with respect to the values of ℓ'_2. This permutation property appears to be a general one for all odd values of ℓ'_1. Also given in the table are the computed mean and standard deviation of these ACC values which are therefore the same for all odd values ℓ'_1. Table 2 shows the computed mean and standard deviation of the ACC values as a function of ν for ν up to 10. The decrease in the standard deviation value as ν increases results from the introduction of more ACC values which are closer to zero.

The next case to consider is when $\nu_1 > \nu_2$. Table 3 shows some example ACC values which illustrate the similarities and differences resulting from the different coefficient values. The common behavior is for the ACC values to start at some initial value (for $\nu_1 = 5$ in this case), then eventually decrease by a factor approaching 2.0 for each integer increase in ν_1. The differences in the ACC values can be seen in the four cases which are shown. The cases were chosen according to the signs of the ACC values for $\nu_1 = 5$ and $\nu_1 = 6$, respectively. Four possibilities (++, +-, - + and - -) are represented here. For $\nu_1 > 6$, no polarity changes are indicated and the decrease in absolute value of the ACC by factors close to 2.0 takes over.

VI. GAUSSIAN FORMS

In this section, Gaussian forms of the decreasing terms of the error statistics are written out. They were obtained by substituting the Gaussian forms of Q_x and Q_{xx} (from (68) and (83)) into the decreasing terms defined in the last section. In the interest of a simplified presentation, a shorthand notation is employed with the following definitions:

$$A = -(\sigma/q)^2/2$$

$$B = \mu/q$$

$$G = \begin{cases} (-1)^S \text{ for TCR} \\ 1 \text{ for TCC} \end{cases}$$

$$H = \begin{cases} 1 \text{ for TCR} \\ 1-2^{\nu} \text{ for TCC} \end{cases} \tag{141}$$

$$V = 1/(\cos(\omega_s \lambda) - 1)$$

$$W = \text{sinc}(\omega_s/2\pi)$$

$$\omega_s = 2\pi s/2^{\nu} .$$

Subscripts are used with G, H, V, W and ω_s whenever more than one coefficient is involved in the formula of the statistic. These have the form

$$G_i = \begin{cases} (-1)^{s_i} \text{ for TCR} \\ 1 \text{ for TCC} \end{cases}$$

$$H_i = \begin{cases} 1 \text{ for TCR} \\ 1-2^{\nu_i} \text{ for TCC} \end{cases} \tag{142}$$

$$V_i = 1/(\cos(\omega_{s_i} \lambda_i) - 1)$$

$$W_i = \text{sinc}(\omega_{s_i}/2\pi)$$

$$\omega_{s_i} = 2\pi\, s_i/2^{\nu_i}$$

for i = 1,2. There is a special term ω_1 used for $\downarrow R_{x_1 \epsilon_2}$ for $|\rho| < 1.0$. In this case

$$\omega_1 = 2\pi\, s_1 .$$

It is assumed that the coefficients have fixed values. Thus, whenever the parameter k′ appears, a specific transformation from k′ to e is implied according to (58). The parameter λ is the coefficient related value obtained from (104).

PROBABILITIES

$$\downarrow P_{\epsilon}(e) = 2^{1-\nu} \sum_{s=1}^{\infty} W \exp(A\omega_s^2) \cos(\omega_s[B-k']) \tag{143}$$

$$\downarrow P_{\epsilon_1\epsilon_2}(e_1,e_2) = 2^{1-\nu_1-\nu_2} \left\{ \sum_{s_1=1}^{\infty} W_1 \exp(A\omega_{s_1}^2) \cos(\omega_{s_1}[B-k'_1]) + \sum_{s_2=1}^{\infty} W_2 \exp(A\omega_{s_2}^2) \cos(\omega_{s_2}[B-k'_2]) \right.$$

$$+ \sum_{s_1=1}^{\infty} \sum_{s_2=1}^{\infty} W_1 W_2 \left[\exp\left\{ A(\omega_{s_1}^2 + 2\rho\,\omega_{s_1}\omega_{s_2} + \omega_{s_2}^2) \right\} \cos(\omega_{s_1}[B-k'_1] + \omega_{s_2}[B-k'_2]) \right.$$

$$\left. \left. + \exp\left\{ A(\omega_{s_1}^2 - 2\rho\omega_{s_1}\omega_{s_2} + \omega_{s_2}^2) \right\} \cos(\omega_{s_1}[B-k'_1] - \omega_{s_2}[B-k'_2]) \right] \right\}. \tag{144}$$

MEAN ERROR

TCR and TCC

$$\downarrow \mu_{\epsilon} = 2^{-M-\nu} \sum_{\substack{s=1 \\ s \not\equiv 0 \bmod 2^{\nu}}}^{\infty} G\,VW \exp(A\omega_s^2) \left[\cos(\omega_s[B + \lambda\,\mathrm{sgna}]) - \cos(\omega_s B) \right]. \tag{145}$$

$R_{\epsilon}(0)$

TCR

$$\downarrow R_{\epsilon}(0) = -2^{-2M-2\nu+1} \sum_{\substack{s=1 \\ s \not\equiv 0 \bmod 2^{\nu}}}^{\infty} (-1)^s\,VW \exp(A\omega_s^2) \cos(\omega_s B). \tag{146}$$

TCC

$$\downarrow R_{\epsilon}(0) = -2^{-2M-2\nu+1} \sum_{\substack{s=1 \\ s \not\equiv 0 \bmod 2^{\nu}}}^{\infty} VW \exp(A\omega_s^2) \left[2^{\nu-1} \cos(\omega_s[B + \lambda\,\mathrm{sgna}]) + (1 - 2^{\nu-1}) \cos(\omega_s B) \right]. \tag{147}$$

$R_{x\epsilon}$ for $\rho = 1.0$

TCR and TCC

$$\downarrow R_{x\epsilon} = 2^{-K-M-\nu} H \sum_{\substack{s=1 \\ s \equiv 0 \bmod 2^{\nu}}}^{\infty} \frac{1}{\omega_s} \exp(A\omega_s^2) \cos(\omega_s/2) \sin(\omega_s B) + 2^{-K-M-\nu} \sum_{\substack{s=1 \\ s \not\equiv 0 \bmod 2^{\nu}}}^{\infty} \exp(A\omega_s^2) \text{ times} \qquad (148)$$

$$\left[\left| W\left(2A\omega_s - \frac{1}{\omega_s}\right) + \frac{1}{\omega_s} \cos(\omega_s/2) \right| VG \left\{ \sin(\omega_s[B + \lambda\, \text{sgna}]) - \sin(\omega_s B) \right\} \right.$$

$$\left. + GBVW \left\{ \cos(\omega_s[B + \lambda\, \text{sgna}]) - \cos(\omega_s B) \right\} \right].$$

$R_{x_1\epsilon_2}$ for $|\rho| < 1.0$

TCR and TCC

$$\downarrow R_{x_1\epsilon_2} = 2^{-K-M-\nu_2} \sum_{\substack{s_2=1 \\ s_2 \not\equiv 0 \bmod 2^{\nu_2}}}^{\infty} G_2 W_2 \exp(A\omega_{s_2}^2) \left[2AV_2 \rho \omega_{s_2} \left\{ \sin(\omega_{s_2}[B + \lambda_2\, \text{sgna}_2]) - \sin(\omega_{s_2} B) \right\} \right.$$

$$\left. + BV_2 \left\{ \cos(\omega_{s_2}[B + \lambda_2\, \text{sgna}_2]) - \cos(\omega_{s_2} B) \right\} \right]$$

$$+ 2^{-K-M-\nu_2} H_2 \sum_{s_1=1}^{\infty} \exp(A\omega_1^2) \frac{(-1)^{s_1}}{\omega_1} \sin(\omega_1 B)$$

$$+ 2^{-K-M-\nu_2} \sum_{s_1=1}^{\infty} \sum_{\substack{s_2=1 \\ s_2 \not\equiv 0 \bmod 2^{\nu_2}}}^{\infty} (-1)^{s_1} G_2 \exp\left\{ A(\omega_1^2 + 2\rho\omega_1 \omega_{s_2} + \omega_{s_2}^2) \right\} \text{ times} \qquad (149)$$

$$\frac{W_2 V_2}{\omega_1} \left[\sin(B[\omega_1 + \omega_{s_2}] + \omega_{s_2} \lambda_2\, \text{sgna}_2) - \sin(B[\omega_1 + \omega_{s_2}]) \right]$$

$$+ 2^{-K-M-\nu_2} \sum_{s_1=1}^{\infty} \sum_{\substack{s_2=1 \\ s_2 \not\equiv 0 \bmod 2^{\nu_2}}}^{\infty} (-1)^{s_1} G_2 \exp\left\{ A(\omega_1^2 - 2\rho\omega_1 \omega_{s_2} + \omega_{s_2}^2) \right\} \text{ times}$$

$$\frac{W_2 V_2}{\omega_1} \left[\sin(B[\omega_1 - \omega_{s_2}] - \omega_{s_2} \lambda_2\, \text{sgna}_2) - \sin(B[\omega_1 - \omega_{s_2}]) \right] .$$

$\mathbf{R_{\epsilon_1\epsilon_2}}$ **for** $\rho = 1.0$

TCR and TCC and $\nu_1 \geqslant \nu_2$

$$\downarrow R_{\epsilon_1\epsilon_2} = 2^{-2M-2\nu_1-\nu_2+1} \sum_{k_1'=0}^{2^{\nu_1}-1} e_1(k_1')\, e_2(k_1') \left[\sum_{s_1=1}^{\infty} W_1 \exp(A\omega_{s_1}^2) \cos(\omega_{s_1}[B - k_1']) \right]. \qquad (150)$$

$\mathbf{R_{\epsilon_1\epsilon_2}}$ **for** $|\rho| < 1.0$

TCR and TCC

$$\downarrow R_{\epsilon_1\epsilon_2} = 2^{-2M-\nu_1-\nu_2-1} H_2 \sum_{\substack{s_1=1 \\ s_1 \not\equiv 0 \bmod 2^{\nu_1}}}^{\infty} G_1 V_1 W_1 \exp(A\omega_{s_1}^2) \left[\cos(\omega_{s_1}[B + \lambda_1 \operatorname{sgna}_1]) - \cos(\omega_{s_1} B) \right]$$

$$+ 2^{-2M-\nu_1-\nu_2-1} H_1 \sum_{\substack{s_2=1 \\ s_2 \not\equiv 0 \bmod 2^{\nu_2}}}^{\infty} G_2 V_2 W_2 \exp(A\omega_{s_2}^2) \left[\cos(\omega_{s_2}[B + \lambda_2 \operatorname{sgna}_2]) - \cos(\omega_{s_2} B) \right]$$

$$+ 2^{-2M-\nu_1-\nu_2-1} \sum_{\substack{s_1=1 \\ s_1 \not\equiv 0 \bmod 2^{\nu_1} \\ s_2 \not\equiv 0 \bmod 2^{\nu_2}}}^{\infty} \sum_{s_2=1}^{\infty} G_1 G_2 V_1 V_2 W_1 W_2 \exp\left\{A(\omega_{s_1}^2 + 2\rho\omega_{s_1}\omega_{s_2} + \omega_{s_2}^2)\right\} \text{times} \qquad (151)$$

$$\left[\cos(B[\omega_{s_1} + \omega_{s_2}] + \omega_{s_1}\lambda_1 \operatorname{sgna}_1 + \omega_{s_2}\lambda_2 \operatorname{sgna}_2) - \cos(B[\omega_{s_1} + \omega_{s_2}] + \omega_{s_1}\lambda_1 \operatorname{sgna}_1) \right.$$

$$\left. - \cos(B[\omega_{s_1} + \omega_{s_2}] + \omega_{s_2}\lambda_2 \operatorname{sgna}_2) + \cos(B[\omega_{s_1} + \omega_{s_2}]) \right]$$

$$+ 2^{-2M-\nu_1-\nu_2-1} \sum_{\substack{s_1=1 \\ s_1 \not\equiv 0 \bmod 2^{\nu_1} \\ s_2 \not\equiv 0 \bmod 2^{\nu_2}}}^{\infty} \sum_{s_2=1}^{\infty} G_1 G_2 V_1 V_2 W_1 W_2 \exp\left\{A(\omega_{s_1}^2 - 2\rho\omega_{s_1}\omega_{s_2} + \omega_{s_2}^2)\right\} \text{ times}$$

$$\left[\cos(B[\omega_{s_1} - \omega_{s_2}] + \omega_{s_1}\lambda_1 \operatorname{sgna}_1 - \omega_{s_2}\lambda_2 \operatorname{sgna}_2) - \cos(B[\omega_{s_1} - \omega_{s_2}] + \omega_{s_1}\lambda_1 \operatorname{sgna}_1) \right.$$

$$\left. - \cos(B[\omega_{s_1} - \omega_{s_2}] - \omega_{s_2}\lambda_2 \operatorname{sgna}_2) + \cos(B[\omega_{s_1} - \omega_{s_2}]) \right] .$$

VII. DISCUSSION

It is appropriate at this point to compare the results derived in this report with the error models presently used in digital filter design. Obviously, simplification of the derived equations are not possible when σ/q is small enough so that the decreasing form of each statistic cannot be ignored. For the purpose of this discussion, σ/q will be assumed large enough so that the decreasing forms are negligible.

Assume a filter structure where the coefficient values are fixed. Associated with each coefficient value is the parameter ν which is the effective word length of the associated multiplication error. The errors can take on each of 2^{ν} discrete values which are uniformly distributed. The probability of occurrence of each value is (almost) equal to $2^{-\nu}$. The mean error, $\vec{\mu}_{\epsilon}$, is

$$\vec{\mu}_{\epsilon} = \begin{cases} 2^{-M-\nu-1} & \text{for TCR} \\ 2^{-M-1}(2^{-\nu}-1) & \text{for TCC} \end{cases} . \tag{152}$$

The error variance, $\vec{C}_{\epsilon}(0)$, is

$$\vec{C}_{\epsilon}(0) = \frac{2^{-2M}}{12}(1-2^{-2\nu}) . \tag{153}$$

An idealized continuous uniformly distributed error ϵ_c is often used in filter design. This error has the following properties:

$$\mu_{\epsilon_c} = \begin{cases} 0 & \text{for TCR} \\ -2^{-M-1} & \text{for TCC} , \end{cases} \tag{154}$$

and

$$C_{\epsilon_c}(0) = \frac{2^{-2M}}{12} . \tag{155}$$

Note that, in comparing (154) with (152) and (155) with (153), the idealized continuous model can only be assumed if ν is large enough. Thus, for example, if the word lengths K, L, M and N are equal, the coefficient values $\pm 1/2$ (or $\nu = 1$) would yield the most difference between the discrete and continuous model. The difference becomes less for coefficient values of $\pm 1/4$ and $\pm 3/4$ (or $\nu = 2$), and so on for larger values of ν.

Other results are the following. The covariance $\vec{C}_{x\epsilon}$ is zero in agreement with the continuous model. This means the multiplier input is uncorrelated with the error. The error covariances are:

$$\vec{C}_{\epsilon}(\eta T) = \begin{cases} \vec{C}_{\epsilon}(0) & \text{for } \eta = 0 \\ 0 & \text{for } \eta \neq 0 \end{cases} \tag{156}$$

and

$$\vec{C}_{\epsilon_1\epsilon_2} = \begin{cases} \vec{C}_{\epsilon_1\epsilon_2} & \text{for } \rho = 1.0 \\ 0 & \text{for } |\rho| < 1.0 \ . \end{cases} \tag{157}$$

These latter results imply that these contributions to the FIR filter variance and output spectrum are white noise in nature.

The consequences of small values of σ/q on the statistics will be explored in a subsequent report. Particular attention will be paid to those ranges of values of σ/q over which the asymptotic model can be assumed.

$\vec{\rho}_{\epsilon_1\epsilon_2}$	$\ell_1' = 1$	$\ell_1' = 3$
1.0000	1	3
0.3548	3, 11	1, 9
0.2023	5, 13	7, 15
0.1202	7, 23	5, 21
0.0616	9, 25	11, 27
–0.0674	15	13
0.2493	17	19
–0.0205	19, 27	17, 25
–0.1730	21, 29	23, 31
–0.8182	31	29
	Values of ℓ_2' which result in the indicated values of $\vec{\rho}_{\epsilon_1\epsilon_2}$	

NOTES:

1. $\nu_1 = \nu_2 = 5$
2. Mean of $\vec{\rho}_{\epsilon_1\epsilon_2} = 0.0909$
   Standard deviation = 0.3566

Table 1. Examples of $\vec{\rho}_{\epsilon_1\epsilon_2}$ for TCR.

ν	Mean of $\vec{\rho}_{\epsilon_1\epsilon_2}$	Standard Deviation
1	1.0	0.0
2	0.6	0.4
3	0.33333	0.4762
4	0.17647	0.4216
5	0.09091	0.3566
6	0.04615	0.2737
7	0.02326	0.2052
8	0.01167	0.1497
9	0.00585	0.1086
10	0.00293	0.0777

Table 2. Computed mean and standard deviation of $\vec{\rho}_{\epsilon_1\epsilon_2}$ for TCR as a function of $\nu_1 = \nu_2 = \nu$.

ν_1	$\ell_1' = 17$ (++)	$\ell_1' = 3$ (+-)	$\ell_1' = 19$ (-+)	$\ell_1' = 15$ (--)
5	0.249267	0.354839	-0.020528	-0.067449
6	0.124588	-0.104067	0.083547	-0.127519
7	0.062288	-0.052029	0.041770	-0.063754
8	0.031143	-0.026014	0.020884	-0.031876
9	0.015572	-0.013007	0.010442	-0.015938
10	0.007786	-0.006503	0.005221	-0.007969

NOTE: $\nu_2 = 5, \ell_2' = 1$

Table 3. Examples of the behavior of $\vec{\rho}_{\epsilon_1\epsilon_2}$ for TCR when $\nu_1 \geqslant \nu_2$.

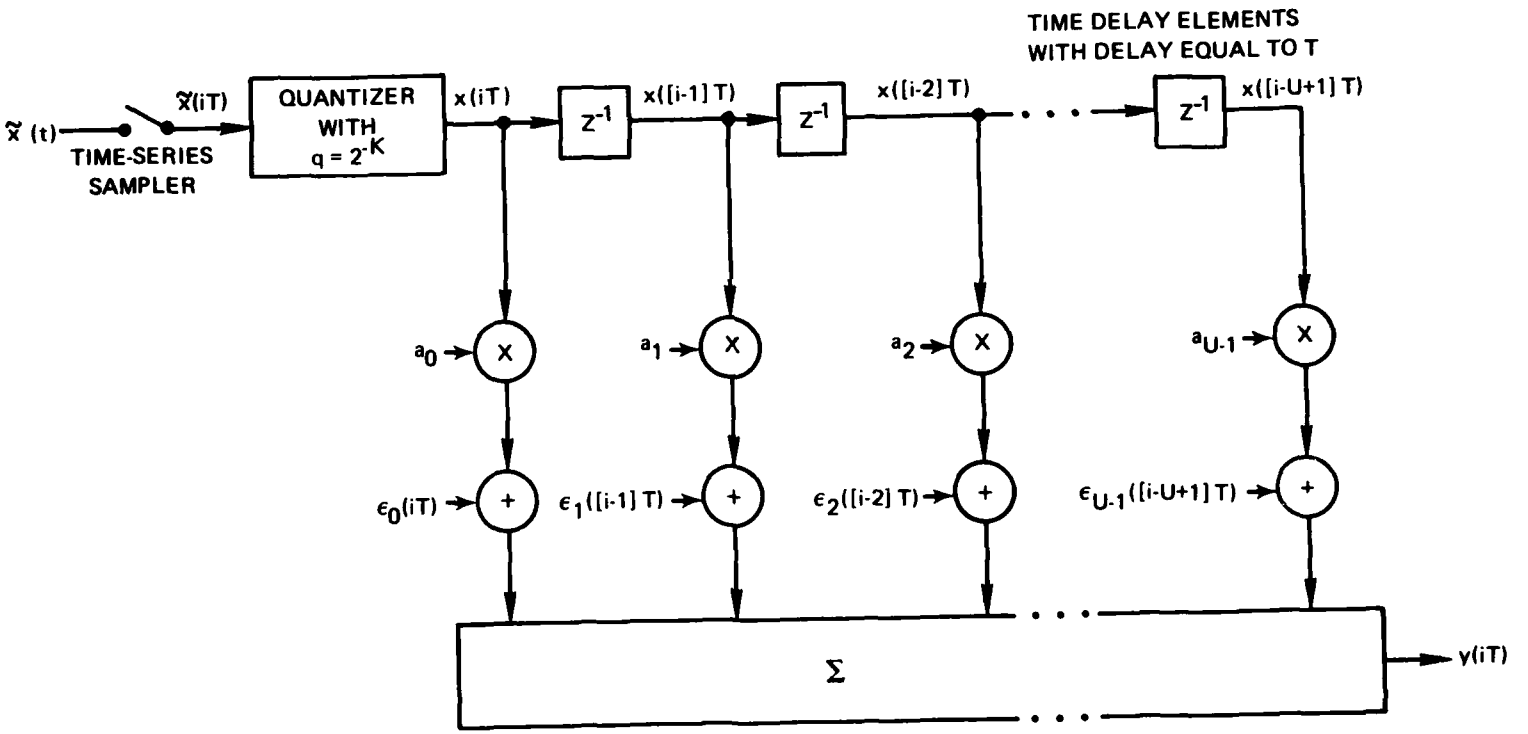


Figure 1. Block diagram representation of an FIR filter which incorporates product quantization errors.

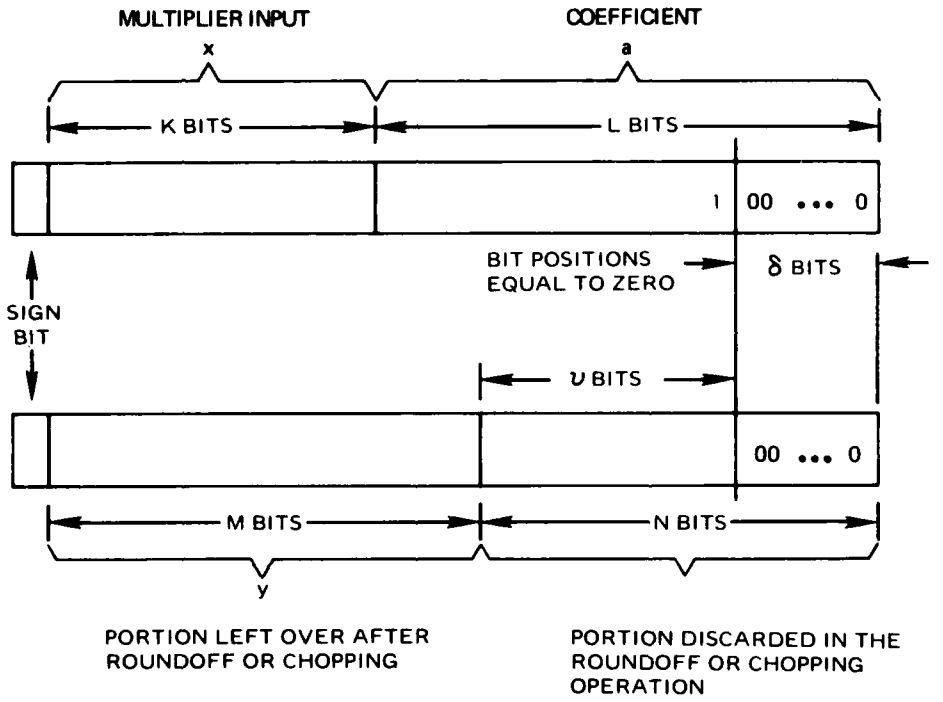
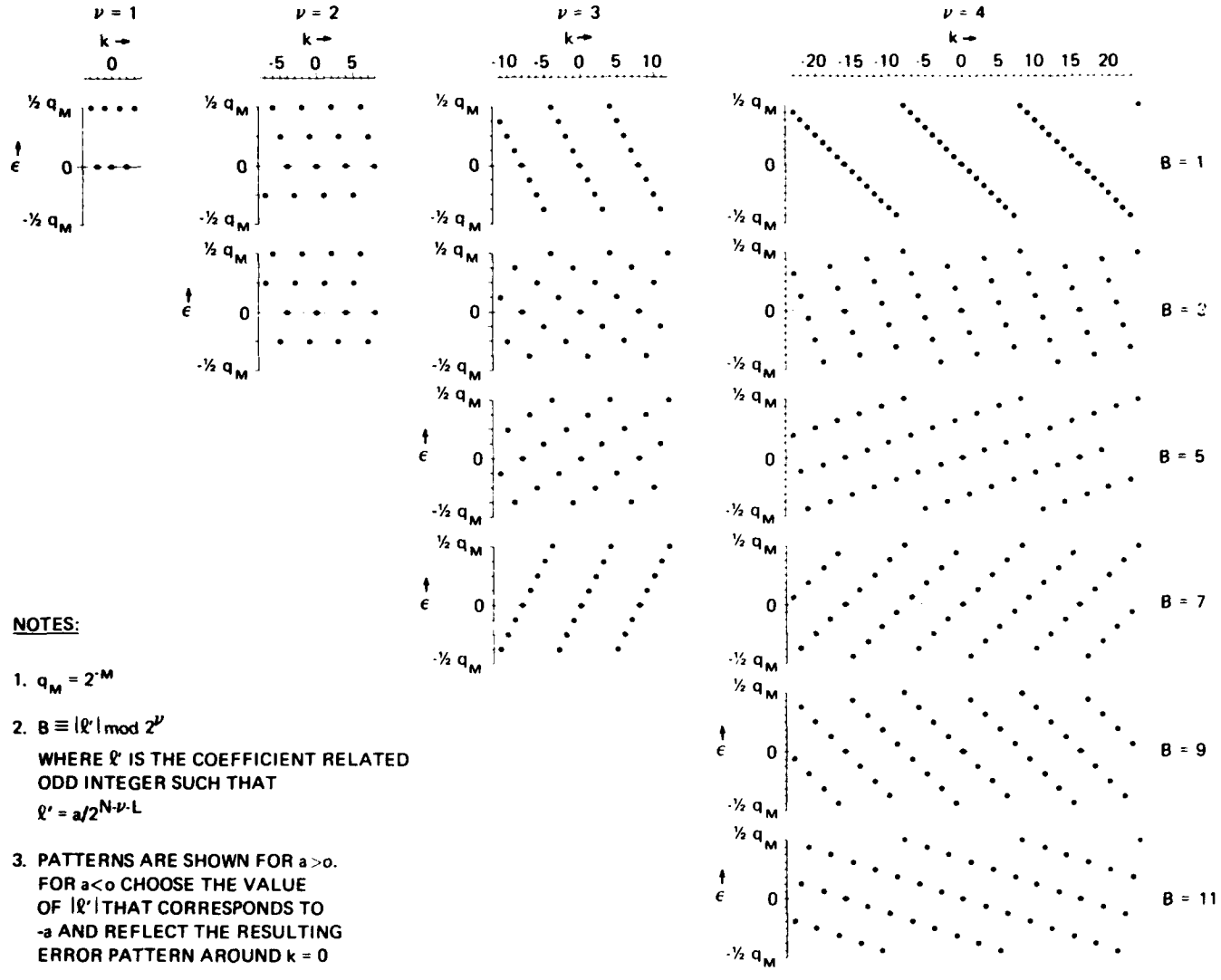


Figure 2. Relationships among the various word lengths.



NOTES:

1. $q_M = 2^{-M}$
2. $B \equiv |\ell'| \bmod 2^{\nu}$
   WHERE ℓ' IS THE COEFFICIENT RELATED ODD INTEGER SUCH THAT
   $\ell' = a/2^{N-\nu-L}$
3. PATTERNS ARE SHOWN FOR a >o. FOR a<o CHOOSE THE VALUE OF $|\ell'|$ THAT CORRESPONDS TO -a AND REFLECT THE RESULTING ERROR PATTERN AROUND k = 0

Figure 3. Graphical example of the mappings from the multiplier input to the error for TCR and positive coefficient values.

VIII. REFERENCES

[1] L. R. Rabiner and B. Gold, *Theory and application of digital signal processing*, Englewood Cliffs, NJ: Prentice-Hall, Inc., 1975, ch. 5, pp. 295–355.

[2] A. V. Oppenheim and R. W. Schafer, *Digital signal processing*, Englewood Cliffs, NJ: Prentice-Hall, Inc., 1975, ch. 9, pp. 404–479.

[3] L. P. Mulcahy, "Digital fixed-point multiplication error structure and some consequences," in *Conf. Rec., 1976 IEEE Conf. Acoust., Speech, Signal Processing*, 1976, pp. 529–532.

[4] S. Bochner, *Harmonic analysis and the theory of probability*, Berkeley and Los Angeles, CA: University of California Press, 1960, p. 32.

[5] A. Papoulis, *Probability, random variables, and stochastic processes*, New York, NY: McGraw-Hill, Inc., 1965, pp. 157 and 209.

[6] P. E. Girard, *Correlated noise effects of structure in digital filters*, Ph.D. dissertation, Naval Postgraduate School, Monterey, CA, June 1974.

[7] S. R. Parker and P. E. Girard, "Correlated noise due to roundoff in fixed point digital filters," *IEEE Trans. Circuit Syst.*, vol. CAS-23, pp. 204–211, April 1976.

APPENDIX A

Evaluation of D_S proceeds as follows. Consider the TCC case first where $a > 0$. Then by using (58) in (101),

$$D_S = \sum_{k'=0}^{2^\nu-1} e_+(k') \exp(-jk'\omega_S)$$

$$= -\sum_{k'=0}^{2^\nu-1} [k'\ell'] \, 2^\nu \exp(-jk'\omega_S) \qquad \text{(A-1)}$$

for $\omega_S = 2\pi s/2^\nu$ where $s = 0, \pm 1, \pm 2, \ldots$, and ℓ' is the coefficient dependent value which is derived from the value of a through (47), (51). Let the variable β be defined by relation

$$\beta \equiv k'\ell' \bmod 2^\nu \qquad \text{(A-2)}$$

which, since ℓ' is an odd positive integer, implies the relation

$$k' \equiv \beta\lambda \bmod 2^\nu \qquad \text{(A-3)}$$

where the odd constant integer $\lambda \in [0, 2^\nu)$ is the unique inverse of ℓ' which satisfies the relation

$$1 \equiv \ell'\lambda \bmod 2^\nu \ . \qquad \text{(A-4)}$$

Since the relationship between k' and β is one-to-one,

$$D_S = -\sum_{\beta=0}^{2^\nu-1} \beta \exp(-j\omega_S[\beta\lambda] \, 2^\nu)$$

$$= -\sum_{\beta=0}^{2^\nu-1} \beta \exp(-j\omega_S\beta\lambda) \ . \qquad \text{(A-5)}$$

Whenever $s \equiv 0 \bmod 2^\nu$ this form simplifies to

$$D_S = -\sum_{\beta=0}^{2^\nu-1} \beta$$

$$= 2^{\nu-1}(1-2^\nu) \ . \qquad \text{(A-6)}$$

This form can be evaluated for other values of s by writing it as

$$D_s = \sum_{\beta=0}^{2^\nu-1} \frac{1}{j\omega_s} \frac{d}{d\lambda} \left\{ \exp(-j\omega_s \beta\lambda) \right\}$$

$$= \frac{1}{j\omega_s} \frac{d}{d\lambda} \left\{ \sum_{\beta=0}^{2^\nu-1} \exp(-j\omega_s \beta\lambda) \right\} . \qquad \text{(A-7)}$$

The summation is over a power series in $\exp(-j\omega_s\lambda)$ which is easily evaluated. By then taking the derivative, the result is

$$D_s = 2^{\nu-1} \frac{\exp(j\omega_s\lambda)-1}{\cos(\omega_s\lambda)-1} \qquad \text{(A-8)}$$

for $s \not\equiv 0 \bmod 2^\nu$.

The computation of D_s for $a < 0$ makes use of the relationship of $e_+(k)$ to $e_-(k)$ as called out in Property 2. Thus, for a_- where $a_- = -a_+$,

$$D_s(a_-) = \sum_{k'=0}^{2^\nu-1} e_-(k') \exp(-jk'\omega_s)$$

$$= 0 \cdot \exp(0) + \sum_{k'=1}^{2^\nu-1} e_+(2^\nu-k') \exp(-jk'\omega_s)$$

$$= \sum_{k'=0}^{2^\nu-1} e_+(k') \exp(jk'\omega_s)$$

$$= D_s^*(a_+) \qquad \text{(A-9)}$$

where * denotes complex conjugate. As a result, the complete expression of D_s for the TCC case, which takes into account the sign of the coefficient value, is written as

$$D_s = \begin{cases} 2^{\nu-1}(1-2^\nu) & \text{for } s \equiv 0 \bmod 2^\nu \\ 2^{\nu-1} \dfrac{\exp((\operatorname{sgn} a)j\omega_s\lambda)-1}{\cos(\omega_s\lambda)-1} , & \text{otherwise} . \end{cases} \qquad \text{(A-10)}$$

The parameter λ is then determined from (A-4) whereby $|\ell'|$ is used in those cases where $a < 0$.

This complex conjugate relationship holds for D_s in the TCR case and F_s in the TCR and TCC cases. Hence, although the bulk of the remaining derivations are carried out for $a > 0$, the final results in each case will be stated so as to include the effect of the sign of the coefficient value.

Evaluation of D_s for the TCR case follows the same procedure as for the TCC case. Substitution of (58) into (10i) for $a > 0$ results in the following form:

$$D_s = {}_1D_s + {}_2D_s \tag{A-11}$$

where

$${}_1D_s = -\sum_{k'=0}^{2^{\nu}-1} [k'\ell' + 2^{\nu-1}]_{2^{\nu}} \exp(-jk'\omega_s) \tag{A-12}$$

and

$${}_2D_s = 2^{\nu-1} \sum_{k'=0}^{2^{\nu}-1} \exp(-jk'\omega_s)$$

$$= \begin{cases} 2^{2\nu-1} & \text{for } s \equiv 0 \bmod 2^{\nu} \\ 0 & \text{otherwise} \end{cases} \tag{A-13}$$

For the evaluation of ${}_1D_s$ let the variable β be defined as

$$\beta \equiv (k'\ell' + 2^{\nu-1}) \bmod 2^{\nu} \tag{A-14}$$

which, since ℓ' is a constant odd integer, implies the relation

$$k' \equiv (\beta\lambda + 2^{\nu-1}) \bmod 2^{\nu} \tag{A-15}$$

where the odd constant integer λ is the unique inverse of ℓ' which satisfies the same relation (A-4) as for the TCC case. With this change in notation

$${}_1D_s = -\sum_{\beta=0}^{2^{\nu}-1} \beta \exp\left(-j\omega_s[\beta\lambda + 2^{\nu-1}]_{2^{\nu}}\right) \tag{A-16 (contd)}$$

$$= - \sum_{\beta=0}^{2^{\nu}-1} \beta \exp(-j\omega_s[\beta\lambda + 2^{\nu-1}])$$

$$= -(-1)^s \sum_{\beta=0}^{2^{\nu}-1} \beta \exp(-j\omega_s \beta\lambda)$$

which is exactly the same form as D_s in the TCC case except for the factor $(-1)^s$. Thus, for TCR, and taking into account the sign of the coefficient value,

$$D_s = \begin{cases} 2^{\nu-1} & \text{for } s \equiv 0 \bmod 2^{\nu} \\ (-1)^s 2^{\nu-1} \dfrac{\exp((\operatorname{sgn} a) j\omega_s \lambda) - 1}{\cos(\omega_s \lambda) - 1} & , \text{ otherwise} \end{cases} \tag{A-17}$$

for $s = 0, \pm 1, \pm 2, \ldots$.

Evaluation of F_s proceeds as follows. Consider the TCC case first where $a > 0$. Then by using (58) in (108)

$$F_s = \sum_{k'=0}^{2^{\nu}-1} e^2(k') \exp(-jk'\omega_s) \tag{A-18}$$

$$= \sum_{k'=0}^{2^{\nu}-1} \left\{ -[k'\ell']_{2^{\nu}} \right\}^2 \exp(-jk'\omega_s)$$

for $\omega_s = 2\pi s/2^{\nu}$ where $s = 0, \pm 1, \pm 2, \ldots$, and ℓ' is derived from the value of a through (47), (51). Let the variables β and λ be defined by the relations (A-2)–(A-4) as for D_s. Then

$$F_s = \sum_{\beta=0}^{2^{\nu}-1} \beta^2 \exp\left(-j\omega_s [\beta\lambda]_{2^{\nu}} \right)$$

$$= \sum_{\beta=0}^{2^{\nu}-1} \beta^2 \exp(-j\omega_s \beta\lambda) \ . \tag{A-19}$$

Whenever $s \equiv 0 \bmod 2^{\nu}$ this form simplifies to

$$F_s = \sum_{\beta=0}^{2^{\nu}-1} \beta^2$$

$$= \frac{1}{6}\left(2 \cdot 2^{3\nu} - 3 \cdot 2^{2\nu} + 2^{\nu}\right) . \tag{A-20}$$

This form can be evaluated for other values of s by writing it as

$$F_s = -\frac{1}{\omega_s^2} \frac{d^2}{d\lambda^2} \left\{ \sum_{\beta=0}^{2^{\nu}-1} \exp(-j\omega_s \beta \lambda) \right\} \tag{A-21}$$

and evaluating the second derivative of the power series in $\exp(-j\omega_s\lambda)$. The result is written for TCC as

$$F_s = \frac{2^{2\nu-1}\left[\exp(j\omega_s\lambda)-1\right]+2^{\nu}}{1-\cos(\omega_s\lambda)} \tag{A-22}$$

for $s \not\equiv 0 \bmod 2^{\nu}$. The complete expression which takes into account the sign of the coefficient value is written as

$$F_s = \begin{cases} \frac{1}{6}\left(2 \cdot 2^{3\nu} - 3 \cdot 2^{2\nu} + 2^{\nu}\right) & \text{for } s \equiv 0 \bmod 2^{\nu} \\ \dfrac{2^{2\nu-1}\left[\exp((\operatorname{sgn} a)j\omega_s\lambda)-1\right]+2^{\nu}}{1-\cos(\omega_s\lambda)} & \text{, otherwise} \end{cases} \tag{A-23}$$

for $s = 0, \pm 1, \pm 2, \ldots$.

Substitution of (58) into (108) for $a > 0$ results in the following form for TCR:

$$F_s = {}_1F_s + {}_2F_s + {}_3F_s \tag{A-24}$$

where

$${}_1F_s = \sum_{k'=0}^{2^{\nu}-1} \left\{ -\left[k'\ell' + 2^{\nu-1}\right]_{2^{\nu}} \right\}^2 \exp(-jk'\omega_s) , \tag{A-25}$$

where

$$ {}_2F_s = -\sum_{k'=0}^{2^\nu-1} 2^\nu [k'\ell' + 2^{\nu-1}]_{2^\nu} \exp(-jk'\omega_s) \ , \tag{A-26} $$

and

$$ {}_3F_s = \sum_{k'=0}^{2^\nu-1} (2^{\nu-1})^2 \exp(-jk'\omega_s) \tag{A-27} $$

$$ = \begin{cases} 2^{3\nu-2} & \text{for } s \equiv 0 \bmod 2^\nu \\ 0 & \text{, otherwise} \ . \end{cases} $$

Let the variable β be defined by the relation (A-14) as for D_s. Then

$$ {}_1F_s = \sum_{\beta=0}^{2^\nu-1} \beta^2 \exp\left(-j\omega_s [\beta\lambda + 2^{\nu-1}]_{2^\nu}\right) $$

$$ = (-1)^s \sum_{\beta=0}^{2^\nu-1} \beta^2 \exp(-j\omega_s \beta\lambda) \tag{A-28} $$

which is exactly the same form as F_s in the TCC case except for the factor $(-1)^s$. Thus

$$ {}_1F_s = \begin{cases} \frac{1}{6}(2\cdot 2^{3\nu} - 3\cdot 2^{2\nu} + 2^\nu) & \text{for } s \equiv 0 \bmod 2^\nu \\ (-1)^s \left\{ \dfrac{2^{2\nu-1}[\exp(j\omega_s\lambda)-1]+2^\nu}{1-\cos(\omega_s\lambda)} \right\} & \text{, otherwise} \ . \end{cases} \tag{A-29} $$

The factor ${}_2F_s$ can be written from (A-12) as

$$ {}_2F_s = 2^\nu \ {}_1D_s $$

$$ = \begin{cases} 2^{2\nu-1}(1-2^\nu) & \text{for } s \equiv 0 \bmod 2^\nu \\ (-1)^s \, 2^{2\nu-1} \dfrac{\exp(j\omega_s\lambda)-1}{\cos(\omega_s\lambda)-1} & \text{, otherwise} \ . \end{cases} \tag{A-30} $$

The final result is

$$F_s = \begin{cases} \frac{1}{12}\left\{2^{3\nu} + 2\cdot 2^{\nu}\right\} & \text{for } s \equiv 0 \bmod 2^{\nu} \\ (-1)^s \dfrac{2^{\nu}}{1-\cos(\omega_s \lambda)}\,, & \text{otherwise} \end{cases} \qquad \text{(A-31)}$$

for $s = 0, \pm 1, \pm 2, \ldots$. Note that this form is real and, hence, there is no need to take into account the sign of the coefficient value.

GLOSSARY

ACRONYMS

ACC	Asymptotic correlation coefficient
A/D	Analog-to-digital
FIR	Finite impulse response
PSF	Poisson summation formula
TC	Two's-complement
TCC	Two's-complement chopping
TCR	Two's-complement roundoff

ABBREVIATIONS

l.s.b.	Least significant bit
p.d.f.	Probability density function
r.v.	Random variable

SYMBOLS

		Equation
C	Auto- and cross-covariances	(8)
E	Denotes expectation or expected value	(5)
I	Indicator function	(27)
K	Multiplier input word length in bits	
L	Multiplication coefficient word length in bits	
M	Multiplier output word length in bits (result of roundoff or chopping)	
N	Number of bits discarded in the roundoff or chopping operation (K,L,M do not include sign bit. See Fig. 2.)	
P	Probability	(64)
Q	Characteristic function or Fourier Transform of a continuous function	(65)
R	Auto- and cross-correlation	(9)
S	Power spectral density	(12)
T	Time interval between consecutive data samples	(1)

U	Number of FIR filter taps	(4)
a	Multiplication coefficient value	(2)
b	Desired or ideal filter output sequence	(1)
e	Integer form of the multiplication error ϵ	(58)
j	$\sqrt{-1}$	(12)
k	Integer form of the filter input sequence x	(46)
ℓ	Integer form of the coefficient a	(47)
q	Quantization step size	(64)
r	Remainder	(35)
u	Integer form of the desired filter output sequence b	(31)
w	Additive noise contribution to filter output sequence y	(1)
x	Filter or multiplier input data sequence	(2)
y	Filter or multiplier output data sequence	(1)
$\tilde{x}$	Analog waveform used for quantizer input	
*	Denotes machine representation (e.g., b*); also, complex conjugate	(28) (A-9)
Δ	Change in a statistic imposed by the multiplication errors	(6)
δ	Number of consecutive zeroes in the l.s.b. positions in the TC representation of the coefficient value	(51)
ϵ	Multiplication error sequence	(3)
λ	Coefficient related odd integer	(104)
μ	Mean value	(4)
ν	Effective word length in bits of the multiplication error	(54)
ρ	Correlation coefficient ($\|\rho\| \leq 1.0$)	
σ	Standard deviation; with no subscript σ denotes $\sigma_{\tilde{x}}$	(5)
ω	Radian frequency	(12)